Research Note 86-100

AD-A178 756

TRAINING FOR RETRIEVAL OF KNOWLEDGE UNDER STRESS THROUGH ALGORITHMIC DECOMPOSITION

Dan Zakay, Colin Kessel, and Lydia Bekhor
Perceptronics

for

Contracting Officer's Representative
Michael Kaplan

BASIC RESEARCH LABORATORY
Milton R. Katz, Director

U. S. Army
Research Institute for the Behavioral and Social Sciences

December 1986

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON Technical Director	L. NEALE COSBY Colonel, IN Commander

Research accomplished under contract
for the Department of the Army

Calspan Corporation

Technical review by

Steve Kronheim

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ARI Research Note 86-100

2. GOVT ACCESSION NO.
AD-A178756

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
TRAINING FOR RETRIEVAL OF KNOWLEDGE UNDER STRESS THROUGH ALGORITHMIC DECOMPOSITION

5. TYPE OF REPORT & PERIOD COVERED
Interim Report
June 85 - May 86

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Dan Zakay, Colin Kessel, and Lydia Bekhor

8. CONTRACT OR GRANT NUMBER(s)
DAJA 45-85-C-0029

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Perceptronics
3, Ha'aretz Street
Tel Aviv 69866, Israel

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q161102B74F

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600

12. REPORT DATE
October 1986

13. NUMBER OF PAGES
99

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
- -

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
- -

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
- -

18. SUPPLEMENTARY NOTES
Michael Kaplan, contracting officer's representative

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Algorithmic Decomposition	Stress
Heuristic	Tutorial
Cognition	Cognitive Style
Probability Estimation	

(document ,

this

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

~~The present~~ study deals with problems of estimation; specifically, the kind performed by military officers under conditions of uncertainity and stress. Two types of estimation were examined in this ~~Research Note~~: the estimation of factual quantities, and the revision of probability in base-rate type problems. Two corrective procedures were also tested: algorithmic decomposition, in which the target estimate is divided into simple and known sub-estimates, which are then combined according to rule, in order to yield the target estimate; and tutorial —) over
(over)

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Research Note 86-100

20. Abstract (continued)

instruction in a technique called Training by Mental Imaging (TbMI).

Results of the testing were that the algorithmic decomposition approach is inadequate for a non-academic military population, since it imposes high mental load, and diverts attention to the creation of the algorithm. The TbMI approach was found to be efficient in training and improving performance, and promoted generalization as well. Thus, it is recommended that the use of mental imaging be further developed, and expanded for use in a computer-aided instruction (CAI) plan.

TABLE OF CONTENTS

SUMMARY

The present study deals with problems of estimation, specifically the kind performed by military officers under conditions of uncertainty and stress. Two types of estimation were examined: estimation of factual quantities, and revision of probability in base-rate type problems. Two corrective procedures were tested. The first procedure was Algorithmic Decomposition, in which the target estimate is divided into simple and known sub-estimates. These are combined, according to a rule, to yield the target estimate. The second procedure was Tutorial, and involved creating mental images and explaining the normative solution of base-rate problems. The Algorithmic Decomposition was tested in Experiments I and II. Subjects had to estimate unknown quantities under both normal and time stress conditions. In Experiment I an algorithm was provided as an aid, while in Experiment II, the subjects were trained to compose their own algorithmic aid. The results showed that, although subjects were able to compose algorithms, this aid was not optimal, and failed to improve estimation. The training method for probability assessment, in base-rate problems, was tested in Experiment III and IV. In Experiment III two aids were tested: algorithm and tutorial. The results showed that both aids were effective. However, only the tutorial resulted in generalization. In Experiment IV The Training by Mental Image (TbMI) method was tested. The results showed that the TbMI method was efficient in training and improving performance, relative to the control group. Generalization was observed for trained groups, under all conditions. It was concluded that the algorithmic decomposition approach was inadequate for a non-academic military population, since it imposed high mental load and diverted attention to the creation of the algorithm. It may have been effective, however, if it were adjusted to the differential cognitive styles of the user populations. Since the incorporation of mental images into the Tutorial aid contributed to better understanding and improved performance, the possibility of applying this method for training for the first type of estimation. It is recommended that the use of mental images should be further developed and expanded to be used in a Computer Aided Instructions (CAI) plan.

TABLE OF CONTENTS

APPENDIX D: Base-rate problems for Experiment III

APPENDIX E: Tutorial and Algorithm (Light Bulb version)

APPENDIX F: Base-rate problems for Experiment IV

APPENDIX G: Training by mental image (TBMI)

SUMMARY

The present study deals with problems of estimation, specifically the kind performed by military officers under conditions of uncertainty and stress. Two types of estimation were examined: estimation of factual quantities, and revision of probability in base-rate type problems. Two corrective procedures were tested. The first procedure was Algorithmic Decomposition, in which the target estimate is divided into simple and known sub-estimates. These are combined, according to a rule, to yield the target estimate. The second procedure was Tutorial, and involved creating mental images and explaining the normative solution of base-rate problems. The Algorithmic Decomposition was tested in Experiments I and II. Subjects had to estimate unknown quantities under both normal and time stress conditions. In Experiment I an algorithm was provided as an aid, while in Experiment II, the subjects were trained to compose their own algorithmic aid. The results showed that, although subjects were able to compose algorithms, this aid was not optimal, and failed to improve estimation. The training method for probability assessment, in base-rate problems, was tested in Experiment III and IV. In Experiment III two aids were tested: algorithm and tutorial. The results showed that both aids were effective. However, only the tutorial resulted in generalization. In Experiment IV The Training by Mental Image (TbMI) method was tested. The results showed that the TbMI method was efficient in training and improving performance, relative to the control group. Generalization was observed for trained groups, under all conditions. It was concluded that the algorithmic decomposition approach was inadequate for a non-academic military population, since it imposed high mental load and diverted attention to the creation of the algorithm. It may have been effective, however, if it were adjusted to the differential cognitive styles of the user populations. Since the incorporation of mental images into the Tutorial aid contributed to better understanding and improved performance, the possibility of applying this method for training for the first type of estimation. It is recommended that the use of mental images should be further developed and expanded to be used in a Computer Aided Instructions (CAI) plan.

GENERAL INTRODUCTION

Modern military decision-making is based, in many cases, on probability assessments and estimation of unknown quantities. These have been extensively studied by philosophers, mathematicians and statisticians as well as psychologists.

Modern cognitive psychology has revealed many cognitive biases and limitations which endanger the rationality and optimality of estimation, assessment and decision making (e.g., Koriat, Lichtenstein & Fischhoff, 1980; Kahneman, Slovic & Tversky 1982; Zakay & Wooler, 1984). Tversky & Kahneman (1972a, 1973, 1974,) showed that contrary to the accepted assumption that people perform such mental operations, in an optimal and rational manner, the principles and logic underlying such judgments are much simpler than those expected according to the normative models. These principles are heuristic rules that lead to judgements that may differ essentially and consistently from those derived by normative principles.

For example, Kahneman & Tversky (1973) showed that subjects use the availability heuristic to assess the probability of an event. That is, the probability of the event is assessed by the ease with which instances or occurrences can be brought to mind (Kahneman, Slovic & Tversky, 1982). If this ease is influenced by the relative frequency of the event, a probability estimation, based on this heuristic, would be a better estimation of the objective probability. This ease, however, is also influenced by factors not related to the relative frequency, such as familiarity, emotional salience and the like. These would cause systematic biases since, for example, a single dramatic event would be more easily remembered and therefore would be judged as more probable.

One of the most intensively investigated heuristics is the representativeness heuristic. (Kahneman & Tversky, 1972a; Tversky & Kahneman, 1982b). This heuristic is used in performing two types of judgments: What is the probability that object A belongs to class B? or what is the probability that event A originates from process B? People tend to rely on the degree to which A is representative of B. When A is highly representative of B, the probability that A originates from B (or B generates A) is judged to be high. This representativeness can be expressed by the degree to which A resembles B, by the statistical relation or causal relation between the two events.

One implication of the representativeness heuristic, is the Base-Rate Fallacy. When asked to assess the probability of occurrence of an event, in light of previous probability information, people tend to neglect the some of the relevant information (the base-rate), which must be taken into account. This is called the base-rate fallacy, and it illustrates the insensitivity of people to the sample domain from which a certain event was selected.

In making estimates of unknown quantities, people often start from an initial value, or starting point, and then make adjustments to produce the final estimate. Different starting points yield different estimates, which are biased toward the anchor which is the initial values (Kahneman & Tversky, 1982).

The commander in the future battle field will have to estimate unknown quantities and assess the probabilities of occurrence of various events. This will be done under conditions of uncertainty, time stress and information load. The stress factor, which is a characteristic of battle field, as well as, many other decision making situations, affects estimating and probability assessing. Technological aids that may be available for use in such situations, will help the expert commander to utilize more information and to make better decisions as compared to the present commander. However, the tasks of estimating quantities, assessing probabilities and evaluating the validity and correctness of data presented to him by these aids, will still be of utmost importance, since the commander will have to evaluate the information presented to him and decide how to use it properly. Therefore it is vital to find ways, or aids, for helping an expert commander in optimally performing these tasks. Since research has shown that strategies, effectively applied to normal conditions, are not transfered to stress conditions (Zakay & Wooler, 1984), such aids will have to be adjusted to stress conditions.

Various such aids were sugested and tested. For example, the algorithmic decomposition technique for estimating unknown quantities (MacGregor, Lichtenstein & Slovic, 1985). As aids in probability assessment, mainly for solving base-rate problem, a number of techniques were suggested, For example, the Subjective Sensitivity Analysis (SSA) (Fischhoff, Slovic & Lichtenstein, 1979; Fischhoff & Bar-Hillel, 1984), the Balanced SSA (BSSA), Isolation Analysis (AI) and Minimal Focusing (FM) (Fischhoff & Bar-Hillel, 1984), and Structuring Base-Rates (Lichtenstein & MacGregor, 1985). These aids were shown to influence people's judgments, however, there was no

contribution to their understanding, or constructively change the way in which people conceputalized the problems (Fischhoff & Bar-Hillel, 1984).

The general objectives of the present study are as follows:

a. To examine the effectiveness of two aiding methods, the algorithmic decomposition technique and Training by Mantal Images (TbMI); and

b. To validate the utility of applying these aiding methods by testing them on military officers, under stress conditions.

The study was carried out in two phases. Phase A focused on the following goals:

a. Testing the effectiveness of the algorithmic decomposition technique, previously investigated by MacGregor, Lichtenstein & Slovic (1985), on Israeli military population under time stress conditions;

b. Development and testing the application of a training method for creating algorithms by the user himself, and testing its effectiveness in performing general estimation tasks under time stress conditions.

Phase B focused on the following goals:

a. Testing the effectiveness of the general training methods in solving base-rate problems, developed by Lichtenstein & MacGregor (198x), on Israeli population.

b. Improving this aiding method by introducing mental imagery and testing its effectiveness on military population, using problems of military content and under time stress conditions.

This report is divided into three parts. In part A the experiments involving estimation of unknown quantity are described and discussed. Part B describes and discusses the experiments involving base-rate problem solving. Part C is a general discussion of the results obtained in the experiments described in part A and B of this report, and general conclusions.

PART A

ALGORITHMIC DECOMPOSITION AS AN AID FOR ESTIMATION

Introduction

In many decision making situation, decisions are based on a set of data concerning different aspects of the situation at hand. These aspects usually involve the values of certain quantities. Some of these values are readily available and easily obtained by reference to various information sources. In many cases, however, the necessary data is unavailable and therefore has to be estimated. This is especially true for battle-field situations, when the availability of information is restricted.

If reliable decisions are to be made, then the estimates must be made as accurately as possible. In addition, in many decision making situations the time factor is very critical, and thus the estimates must also be made as quickly as possible.

When considering the cognitive strategies people adopt in making estimates, it is realized that such strategies are usually ineffective and lead to inadequate estimates. For example, one approach is to consider one's knowledge of the quantity being estimated, and intuitively guess an estimate that seems reasonable in light of whatever knowledge comes to mind (MacGregor, Lichtenstein & Slovic (1985). Another approach is to start from an initial value, and then adjust it to yield the final estimate (Kahneman, Slovic & Tversky, 1982). Estimates performed according to such approaches, are highly influence by irrelevant and biasing factors. For example, the initial values may become anchors, therefore the estimates are biased toward them (Tversky & Kahneman, 1974). In addition, recency, salience and emotional factors influence the ease with which information is retrieved from memory. This may affect the adjustments of the initial values (Nisbett & Ross, 1980). Since the resulting estimates are determined by subjective factors, they do not represent the state of the world. Decisions based on such intuitive estimates may be erroneous.

The inefficiency of the intuitive strategies, and the frequent demand for accurate estimates, which have to be made as quickly as possible, make way to development and application

of aids, specifying alternative approaches to making estimates.

The algorithmic decomposition is one such approach. It involves the division of an estimation question to a series of sub-questions, the answers for which are more accurate, easily obtained and of which one is more likely to have available knowledge. The answers to the sub-questions can then be combined, according to a rule, to yield the answer to the original estimation question. The resulting estimate would be more accurate than a direct estimate. The approach of analysis and decomposition is based on the concept of structuring information in accordance with knowledge organization in human memory, in order to obtain and utilize information from various external sources, as well as retrieval from memory. It is assumed that an intuitive wholistic strategy of estimation, which incorporates less knowledge, creates a "vacuum", into which the heuristics are introduced. By using the algorithmic decomposition approach, this "vacuum" can be filled with knowledge, and reduce out the biases.

The utility of the algorithmic decomposition approach, was tested by MacGregor, Lichtenstein, & Slovic, (1985). Their subjects had to estimate the answers to various questions concerning unknown quantities. This was done under five aiding conditions, constituting different structuring levels, as follows:

a. Full Algorithm. In this condition, each question was decomposed into a complete algorithm. Subjects had to estimate the answers to the sub-question and then combine the sub-estimates according to the provided arithmetic rules that were provided.

b. Partial Algorithm. This condition, was similar to the full algorithm, however, the arithmetic rules were not provided.

c. List & Estimate. In this condition, subjects, first, listed components or factors that they believed were relevant to estimating the target quantity; they then estimated each of the components they had listed; after completing these tasks, the subjects made the target estimates.

d. List. This condition was similar to list & estimate, however, the subjects were not requested to make estimates of the factors they have listed.

e. Unaided. In this condition, no aiding was provided to subjects in making the target estimates.

The results of this experiment showed that accuracy and consistency was improved with increasing the structuring level. That is,the partial algorithm and full algorithm condition produced more accurate and consistent estimates then the list & estimate, list and unaided conditions. Similarly, the list condition led to more accurate estimates than the unaided condition, however, the list & estimate condition did not help the subjects focus more directly on the magnitude of the value they had to estimate.

In their discussion, MacGregor, Lichtenstein, & Slovic (1985), pointed out that, although their results showed that people can perform estimates according to specified algorithms, in evaluating the quality of such an approach one has to take into consideration the representation of the estimation questions. Estimation question can be represented in several ways, each of them may influence the subject's performance in a different manner. A representation which is effective for some people may be biasing and misleading for others. This means that to be a useful aid, an algorithm must be compatible with the specific needs of each user. In addition, in everyday, and especially, battle-field decision making situations, algorithms are not provided, instead they have to be produced by the decision maker.

Battle-field decision making, is performed under time stress conditions, and is known to be characterized by unique cognitive processes. Ben Zur and Breznitz (1981) found that under high time pressure, a filtration mechanism was activated, that is, "Information which was perceived by the individual as most important was processed first, and then processing was continued until time was up" (p. 102). Research has shown that framing is not transferred to stress conditions. In addition, Zakay, (1985) found that time pressure would lead to more frequent use of non-compensatory strategies.

In view of the above, the algorithmic decomposition may become an effective aid, if taught and used as a mental model of estimation. This would enable people to compose their own individual algorithms, which would be compatible with their own mental processes and needs. Since, the utility of this approach may be dependent on the user, and can be affected by stress conditions (similarly to other cognitive aids), it is

important to test its effectiveness on various populations, mainly military populations, using stress conditions.

The purpose of the present study is to evaluate the algorithmic decomposition, as a method for estimating quantities, by military officers under normal and time stress conditions. An additional goal is to develop and validate a method, based of the algorithmic decomposition approach, for training military officers in composing and using their own algorithms when estimating quantities, under normal and time stress conditions.

Experiment I

Experiment I was designed to test the effectiveness of algorithmic decomposition, as an aid in estimating unknown quantities, on military population. Experiment I is a partial replication of the study carried out by MacGregor, Lichtenstein, & Slovic, (1985). However, of the five structuring levels used in the original experiment, only the full algorithm, which was found to be the most effective, and the unaided control levels were used in the present experiment.

Method

Subjects. Seventy one IDF junior officers participated in experiment I. The subjects have had secondary education.

Estimation tasks. All the subjects were asked to perform 8 different estimation tasks. The questions were of the type "How much food (in Kg.) does one person eat during his entire life". The corract answers to the questions varied in magnitude from 27,040,000,000 to 100. The estimation question and the correct answers are show in Appendix A. The questions were based on quantities contained in statistical almanacs. A pilot study, revealed that the subjects were unlikely to know the exact quantities to which the estimation questions related, however, the they did have some relevant knowledge, on which the sub-estimates could be based.

Experimental design. Two independent variables were manipulated in this experiment: aid type, and time restriction.

Aid type: In the aided condition a full algorithm was provided as an aid in answering each estimation question. Under this condition subjects had to estimate each component of

the algorithm and then combine the estimates according to the arithmetic rule provided (this condition is denoted AL). For example:

<u>How many cigarettes are consumed in Israel in a year</u>?

a. What is the population of Israel?

b. What proportion of the population smokes?

c. What is the number of smokers in Israel? [multiply (a) by (b)].

d. How many cigarettes does the average smoker consume per day?

e. How many cigarettes are consumed in Israel per day? [multiply the answer to (c) by (d)].

f. How many days are there in a year?

g. How many cigarettes are consumed in Israel in a year? [multiply (e) by (f)].

In the unaided condition, which served as the control condition, no aid was provided. Under this condition the target quantities had to be estimated directly (this condition is denoted C). Each target quantities used in the study was estimated directly.

Time restriction: There were two time restriction conditions. In the unlimited time condition, subjects had to make the estimates without any time restriction (this condition is denoted NTL). In the limited time condition subjects had to answer each estimation problem within 2 minutes (this condition is denoted TL). The 2 minutes limit was determined by running the unlimited time AL group first, and measuring the time required to make each estimation. The minimum time was selected as the time limit for the time restricted groups.

The experimental design with two independent variables, which make up 4 different groups, is presented in Table 1.

Table 1: The Experimental Design of Experiments I and II

AID / TIME LIMIT	AIDED	UNAIDED
LIMITED		
UNLIMITED		

The dependent variables. Two dependent variables were measured in this experiment: subjects' estimates and subjective mental load.

Subjects' estimates. Subjects' responses for each estimation task were recorded. The measure for accuracy of estimation was computed as follows:

$$V = \frac{|X-Y|}{Y}$$

Where V is the measure of accuracy, X is the subject's estimation and Y is the correct answer. This transformation was selected to enable direct comparisons between different estimates, regardless of their actual magnitude. Absolute value were used, since the direction of the error is irrelevant.

Subjective mental load measures. Subjective ratings of subjective mental load were used to measure the subjects' subjective mental load in the various experimental conditions. Subjects rated the extent of difficulty, mental effort, fatigue, frustration, and time limit they have encountered while working on the problems. The rating were made on a seven point scale ranging from "very easy", "little effort", "not tiring at all", "not frustrating at all" and "enough time" (1) to "very difficult", "a lot of effort", "very tiring", "very frustrating" and "not enough time". These were later used for obtaining a measure of subjective mental load. The scales (shown in Appendix B) were adopted from Vidulich & Tsang, (1985).

Procedure. The subjects were randomly assigned to four groups of 15 to 20 officers. Each group of respondents was run separately in small classrooms at an army base. The instructions preceding each session, indicated that the purpose of the study was to examine the ways in which military decision makers solve various problems. It was emphasized that participation in the experiment was anonymous, and that subjects' performance would not affect their career.

The experimental sessions were the same for all the groups. Subjects were given a brief introduction to the study and then filled in a standard form containing details such as age, sex, months of service, current job, command experience, education and the like. The subjects then received the eight estimation problems, and answered them under the experimental conditions,to which they were assigned. During the sessions, pocket calculators were available for use, in order to avoid arithmetic errors which might affect the estimates. Subjects in the TL groups were not told at the beginning of the experimental session, that they would be time restricted later. In this condition all the subjects in a TL groups were asked to make each estimate within a fixed time interval, whose starting and ending points were signaled by the experimenter. After completing the above, the respondents completed the subjective mental load questionnaire.

Results

Accuracy of Estimation

The means and standard deviations (s.d.) of the accuracy measures are presented in Table 2 for each estimation question.

Table 2: Means and Standard Deviations of Accuracy Measures (Upper no.= mean; Lower no.= s.d.)

QUESTION	V1		V2		V3		V4		V5		V6		V7		V8	
AID \ TIME	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL
AIDED	.85	.90	.96	1.62	.83	.81	1.36	.50	9.67	6.15	3.03	10.86	.99	.83	7.08	5.03
	1.53	4.04	.78	2.18	.28	.22	3.84	.40	27.49	15.30	5.65	28.89	1.03	.36	27.12	4.97
UNAIDED	.73	.69	.58	.86	.97	.96	.74	.69	1.04	1.09	.54	.69	1.28	.80	.99	.73
	1.80	1.37	.27	.95	.04	.08	.96	1.23	.89	1.02	.73	.55	2.11	.53	1.16	.26

each estimation question, are shown in Table 3. No interaction effects were obtained.

Table 3: Summary Table of ANOVA of Accuracy Measures for each Estimation Question. (No. indicate F(1,67) values)

QUESTION / VARIABLE	V1	V2	V3	V4	V5	V6	V7	V8
AID TYPE	1.140	*4.255	**11.121	.164	3.274	*3.991	.217	2.113
TIME LIMIT	.488	2.928	.120	.816	.220	1.611	1.348	.027

* p<.05
** p<.01

Table 3 shows that the mean accuracy measures are higher for the AL groups than for the C groups, for the majority of questions (e.g., V1, V2, V4, V5, V6, and V8). The opposite is observed only for questions V3 and V7. The mean accuracy measures are higher for the NTL groups than for the TL groups for questions V1, V2, V3, and V6. The opposite is observed for questions V4, V5, V7, and V8.

Subjective mental load

Difficulty. The difficulty mean ratings and s.d. are presented in Table 4.

Table 4: Means and Standard Deviations of Difficulty Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	3.14	4.75	4.09
	1.16	1.65	1.80
LIMITED	3.75	4.07	3.88
	1.66	1.64	1.64
TOTAL	3.48	4.47	3.99
	1.64	1.66	1.17

Table 4 indicates that the mean difficulty rating across all groups is 3.99. The mean ratings are higher for the C groups (4.47, s.d.=1.66) than for the AL groups (3.48, s.d.=1.64). The mean ratings are higher for the NTL groups (4.09, s.d.=1.80) than for the TL groups (3.88, s.d.=1.64).

An analysis of variance on these data indicated significant effect of aid ($F(1,63)=5.77$, $p<.05$).

Mental effort. The mental effort mean ratings and s.d. are shown in Table 5.

Table 5: Means and Standard Deviations of Mental Effort Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	3.29 1.68	4.40 1.93	3.94 1.87
LIMITED	3.68 1.45	4.07 1.44	3.85 1.44
TOTAL	3.52 1.54	4.26 1.73	3.90 1.66

The data in Table 5 indicate that the mean mental effort rating across all groups is 3.90. The mean ratings are higher for the C groups (4.26, s.d.=1.73) than for the AL groups (3.52, s.d.=1.54). The mean ratings are higher for the NTL groups (3.94, s.d.=1.87) than for the TL groups (3.85, s.d.=1.44).

An analysis of variance on these data showed no significant main effects.

Fatigue. The fatigue mean ratings and s.d. are shown in Table 6. The data in Table 6 show that the mean fatigue rating across all groups is 3.55. The mean ratings are higher for the AL groups (4.06, s.d.=1.75) than for the C groups (3.03, s.d.=1.87). The mean ratings are higher for the NTL groups (3.84, s.d.=1.86) than for the TL groups (3.27, s.d.=1.84).

Table 6: Means and Standard Deviations of Fatigue Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	4.43 3.93	3.39 1.93	3.84 1.86
LIMITED	3.79 1.81	2.57 1.70	3.27 1.84
TOTAL	4.06 1.75	3.03 1.87	3.55 1.86

An analysis of variance on these data indicated significant main effect for aid type ($F(1,61)=6.32$, $p<.05$).

Frustration. The frustration mean ratings and s.d. are shown in Table 7.

Table 7: Means and Standard Deviations of Frustration Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	3.93 2.02	4.17 1.87	4.06 1.91
LIMITED	4.26 2.05	4.00 1.75	4.15 1.91
TOTAL	4.12 2.01	4.09 1.80	4.11 1.90

In Table 7 it can be seen that the mean frustration rating across all groups is 4.11. The mean ratings are higher for the AL groups (4.12, s.d.=2.01) than for the C groups (4.09, s.d.=1.80). The mean ratings are higher for the TL groups (4.15, s.d.=1.91) than for the NTL groups (4.06, s.d.=1.91).

An analysis of variance on these data failed to reach significance.

Subjective time stress. The time stress mean ratings and s.d. are shown in Table 8.

Table 8: Means and Standard Deviations of Time Stress Ratings (Upper No.= Mean; Lower No.=S.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	1.64 1.39	1.39 .83	1.50 1.09
LIMITED	2.68 2.68	3.71 1.94	3.12 2.01
TOTAL	2.24 1.82	2.41 1.82	2.32 1.81

The data in table 8 indicate that the mean time stress rating across all groups is 2.32. The mean ratings are higher for the C groups (2.41, s.d.=1.82) than for the AL groups (2.24, s.d.=1.82). The mean ratings are higher for the TL groups (3.12, s.d.=2.01) than for the NTL groups (1.50, s.d.=1.09).

An analysis of variance on these data indicated significant main effect for time restriction ($F(1,61)=17.23$, $p<.01$).

Subjective mental load. The computed values of subjective mental load are shown in Table 9.

Table 9: Subjective Mental Load Values

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	3.33 1.30	3.62 1.21	3.49 1.24
LIMITED	3.63 1.27	3.69 .88	3.65 1.09
TOTAL	3.50 1.28	3.65 1.06	3.57 1.17

The data in Table 9 show that the mean subjective mental load, across all groups, is 3.57. These values are higher for the C groups (3.65) than for the AL groups (3.50). The values are higher for the NTL groups (3.65) than for the TL groups (3.49).

An analysis of variance on these data failed to reach significance.

Discussion

The results of Experiment I show that the algorithmic decomposition aid, for estimation of unknown quantities, did not lead to more accurate estimates, but rather, it caused enlargement of the errors. That is, the difference between the subjects' estimates and the correct answers is usually larger, when performed with the algorithmic aid, than without any aid.

These results are not in line with those found by MacGregor, Lichtenstein & Slovic, 1985. Their results indicated of more accurate estimates, when the algorithmic decomposition aid was provided. The reason for these contradictory findings, may lie in the difference between the populations tested in each experiment. The subjects, participating in MacGregor, Lichtenstein & Slovic's experiment were university students. For members of such a population, strategies such as the algorithmic decomposition, may be compatible with their own way of thinking, which was acquired as a habit, and became more intuitive. On the other hand, the population tested in Experiment I, was IDF junior officers, whose formal education was non-academic. These subjects are likely to have different and unique patterns of thinking, adapted to their usual tasks and acquired while performing them. These intuitive patterns may be quite different than those imposed by the algorithmic decomposition aid, thus, leading to inaccurate estimates.

Apart from the adequacy or inadequacy of the algorithmic approach, one should consider the fact that estimation errors, occurring in making the sub-estimates, influence the odds for, and the size of error in the target estimate, relative to wholistic direct estimates. For example, suppose a user exhibits a tendency for over estimation, therefore over estimating each component of the algorithm. The accumulated effect of this over estimation would lead to target estimates

which are much larger then estimates performed directly.

In addition to the problems caused by the algorithmic decomposition, as a strategy of estimation, the content of the algorithms (e.g., the specific sub-estimates, The number of sub-estimates, estimation of proportion Vs. integers, etc.) may influence performance. However, this aspect may be more dependent on the cognitive style of the individual user, than on the characteristics of the population in which he/she is a member.

This suggests that the aiding approach, and its content have to be adapted to the specific user population. This is especially important when dealing with military population and should take into account the unique conditions under which it operates.

The subjective mental load measures, obtained here, seem to support these interpretations. Subjects in the aided groups reported higher difficulty and mental effort than those in the unaided groups. Subjects in the time limited groups, also reported higher difficulty and mental effort than those in the unlimited time groups. This may indicate, that wholistic direct estimation cause high subjective mental load that can be reduced by aiding. The higher degree of fatigue and frustration, however, reported by the aided subjects, may be a result of the incompatibility of the specific aid tested.

Experiment II

Experiment I examined the effectiveness of the algorithms, when it is fully provided, as an aid, by the experimenter. Since in real life an algorithm must be composed by the user, a method was developed in order to train people in creating their own algorithms. Experiment II was design to test the effectiveness of such a training method, on military population.

Method

Subjects. Seventy one IDF Junior officers participated in Experiment II. The subjects have had secondary education.

Estimation tasks. The estimation questions employed in Experiment II were the same as those used in Experiment I.

Experiment II were the same as those used in Experiment I.

Experimental design and procedure. The experimental design and procedure in Experiment II was similar to those of experiment I. However, no algorithms were provided, instead, before answering the questions, the aided subjects had to read a detailed tutorial, describing the algorithmic decomposition approach and why it should be used, The tutorial, shown in Appendix C, also explained how to create an algorithm, and contained two training estimation questions. No training was provided for the control groups.

Results

Accuracy of Estimation

The mean and s.d. of the accuracy measures are presented in table 10 for each estimation question.

Table 10: Means and Standard Deviations of Accuracy Measures (Upper no.= mean; Lower no.= s.d.)

QUESTION	V1		V2		V3		V4		V5		V6		V7		V8	
AID \ TIME	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL	TL	NTL
AIDED	1.85	.62	1.74	1.17	1.00	.89	5.59	1.07	8.33	2.48	3.12	3.31	.78	.73	3.94	1.43
	2.83	.52	2.06	1.70	.16	.25	20.58	1.05	16.87	3.00	10.53	9.30	.33	.25	13.30	3.05
UNAIDED	.69	.73	1.38	1.01	.97	.99	4.45	4.92	3.92	6.12	1.14	8.55	.75	.77	.48	.73
	.42	.69	1.97	1.79	.19	.03	.41	10.99	3.30	11.71	.57	33.12	.18	.28	.31	.52

The main effects, obtained in analyses of variance, for neach estimation question, are shown in Table 11. No main or interaction effects were obtained.

Table 11: Summary Table of ANOVA of Accuracy Measures for each Estimation Question. (No. indic ate F(1,67) values)

QUESTION / VARIABLE	V1	V2	V3	V4	V5	V6	V7	V8
AID TYPE	2.366	.344	1.506	.027	.610	.155	.006	1.725
TIME LIMIT	3.366	1.166	1.932	.000	1.204	.718	.045	.527

Table 11 shows that the mean accuracy measures are higher for the aided groups than for the unaided groups for questions V1, V2, V4, V5, and V8. The opposite is observed for questions V3 and V6. The mean accuracy measures are the same for the aided and unaided groups for question V7. The mean accuracy measures are higher for the unlimited time groups than for the limited time groups for questions V4, and V6. The opposite is observed for questions V2, V3, V5, V7, and V8.

Subjective mental load

Difficulty. The difficulty mean ratings are presented in Table 12.

Table 12: Means and Standard Deviations of Difficulty Ratings (Upper No.= Mean; Lower No.=S.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	3.81 1.28	4.40 1.96	4.14 1.69
LIMITED	3.39 1.28	5.15 1.68	4.50 1.16
TOTAL	3.87 1.26	4.70 1.86	4.32 1.64

The data in table 12 indicate that the mean rating, across all groups, is 4.32. The mean ratings are higher for the unaided groups (4.70, s.d.=1.86) than for the aided groups (3.87, s.d.=1.26). The mean ratings are higher for the limited time groups (4.50, s.d.=1.16) than for the unlimited time groups (4.14, s.d.=1.69).

An analysis of variance on these data indicated significant effect of aid ($F(1,60)=4.63$, $p<.05$).

Mental effort. The mental effort mean ratings are shown in Table 13. The data in Table 13 show that the mean rating, across all groups, is 4.30. The mean ratings are higher for the aided groups (4.29, s.d.=1.40) than for the unaided groups (4.27, s.d.=1.91). The mean ratings are higher for the limited time groups (4.46, s.d.=1.84) than for the unlimited time groups (4.14, s.d.=1.55).

Table 13: Means and Standard Deviations of Mental Effort Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	4.38 1.36	3.95 1.70	4.14 1.55
LIMITED	4.20 1.47	4.77 2.21	4.46 1.84
TOTAL	4.29 1.40	4.27 1.91	4.30 1.67

An analysis of variance on these data showed no significant main effects.

Fatigue. The fatigue mean ratings are shown in Table 14.

Table 14: Means and Standard Deviations of Fatigue Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	3.53 1.36	3.28 1.51	3.39 1.44
LIMITED	3.27 1.58	2.86 1.29	3.07 1.44
TOTAL	3.40 1.46	3.06 1.41	3.24 1.44

In Table 14 it can be seen that the mean rating, across all groups, is 3.24. The mean ratings are higher for the aided groups (3.40, s.d.=1.46) than for the unaided groups (3.06 s.d.=1.41). The mean ratings are higher for the unlimited time groups (3.39, s.d.=1.44) than for the limited time groups (3.07, s.d.=1.44).

An analysis of variance on these data showed no significant main effects.

Frustration. The frustration mean ratings are shown in Table 15.

Table 15: Means and Standard Deviations of Frustation Ratings (Upper No.= Mean; Lower No.=S.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	3.27 1.75	4.78 2.24	4.09 2.07
LIMITED	3.53 2.00	4.21 2.08	3.86 2.03
TOTAL	3.40 1.84	4.53 2.15	3.98 2.04

The data in Table 15 indicate that the mean rating, across all groups, is 3.98. The mean ratings are higher for the unaided groups (4.53, s.d.=2.15) than for the aided groups (3.40, s.d.=1.84). The mean ratings are higher for the unlimited time groups (4.09, s.d.=2.07) than for the limited time groups (3.86, s.d.=2.03).

An analysis of variance on these data indicated significant main effect of aid ($F(1,58)=4.88$, $p<.05$).

Subjective time stress. The mean time stress ratings are presented in Table 16. The data in Table 16 indicate that the mean rating, across all groups, is 2.66. The mean ratings are higher for the aided groups (2.90, s.d.=1.99) than for the unaided groups (2.44, s.d.=1.81). The mean ratings are higher for the limited time groups (3.31, s.d.=1.91) than for the unlimited time groups (2.90, s.d.=172).

Table 16: Means and Standard Deviations of Time Stress Ratings (Upper No.= Mean; Lower No.=S.d.)

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	1.87	2.28	2.09
	1.55	1.87	1.72
LIMITED	3.93	2.64	3.31
	1.87	1.78	1.91
TOTAL	2.90	2.44	2.66
	1.99	1.81	1.90

An analysis of variance on these data indicated significant main effect for time restriction ($F(1,58)=6.94$, $p<.05$).

Subjective mental load. The computed measures for subjective mental load are shown in Table 17.

Table 17: Subjective Mental Load Values

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	3.37	3.72	3.56
	1.04	1.40	1.25
LIMITED	3.77	3.81	3.79
	1.17	1.01	1.08
TOTAL	3.57	3.76	3.67
	1.11	1.24	1.18

The data in Table 17 show that the subjective mental load measure, across all groups is 3.67. These measures are higher for the unaided groups (3.76,) than for the aided groups (3.57). The mean ratings are higher for the limited time groups (3.79) than for the unlimited time groups (3.56).

An analysis of variance on these data failed to reach significance.

Discussion of Part A

Content analysis of the questionnaires, used in Experiment II, indicated that the subjects did learn to compose algorithms and were able to apply them successfully. However, the results of Experiment II showed that the training method for building algorithms was not effective. It did not lead to more accurate estimates, but caused enlargement of the estimation errors. That is, the difference between the subjects' estimates and the correct answers was usually larger, when performed with the algorithmic aid, than without any aid. These results are in line with those found in Experiment I. Therefore, it is likely that the reason for the subjects' poor performance, is the inadequacy of the algorithmic decomposition approach. As shown in Experiment I, this is due to the incompatibility of this approach with the characteristics of the military population, used in these experiments, and to the accumulated effects of biased sub-estimation.

When considering the subjective mental load measures, the results showed that the untrained subjects found their task to be more difficult and frustrating than the trained subjects. This indicates the need for aiding. The higher degree of mental effort and fatigue, reported by the trained subjects, however, suggests that the actual creation of an algorithm may be highly demanding, and may divert the users' attention from the estimation itself.

Both Experiments I and II suggest that in developing an aid or training method, based on the algorithmic approach, one should take into account the unique characteristics of the target population, and adjust the aid accordingly. Such an aid would be more compatible with the thinking patterns and cognitive style of the target population. Only after the aid and training method, are adapted to the population, its cognitive style and thinking patterns, an efficient aiding method can be introduced.

PART B

TRAINING FOR OVERCOMING THE BASE-RATE FALLACY

INTRODUCTION

In problem solving and decision making, people tend to ignore some of the available information even though it is relevant. Base-rate problems are inference problems containing base-rate information about a certain phenomenon (prior odds), information about the degree of accuracy of a given diagnostic device or method (usually referred to as the diagnostic or specific information), and a question concerning the probability of a particular event. Base-rate fallacy is the tendency to neglect the Base-rate information when attempting to solve such problems.

The most common example of the base-rate problem is the Cab Driver problem (Tversky & Kahneman, 1972a):

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
>
> a. 85% of the cabs in the city are Green and 15% are Blue.
>
> b. A witness identified the cab as Blue.
>
> The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green?

Research has shown that when presented with this problem or variations of it, people tend to consider only the diagnostic information (e.g., Kahneman & Tversky,1973; Lyon & Slovic, 1976). Therefore the answer, usually given to the above problem is 80%.

The normative statistical model applicable in solving such problems is Bayes' rule or theorem. It is useful in computing probabilities of various hypothesis which have resulted in a

given event (Beyth-Marom & Fischhoff, 1983). Bayes' rule maintains that in re-evaluating the state of the world, one should consider both the new evidence and previous knowledge.

This rule is formulated as follows:

$$P(E1/A) = \frac{P(E1) * P(A/E1)}{P(E1) * P(A/E1) + P(E2) * P(A/E2)}$$

Where P(E1) and P(E2) are the possible states of the world, and P(A) is the new evidence.

Extensive research was done in order to find the conditions under which base-rate information is neglected or used. Some reseachers argued that the biased responses of their subjects stemmed from the content of the problem story. Hammerton (1973) asked subjects to solve base-rate problems in which the diagnostic information referred to the degree of accuracy of a medical diagnostic test. The results showed that subject's judgment were dominated by the diagnostic information. Hammerton argued that the reason for this result was that the subjects had "rigid prior expectation" that such tests are infallible, and therefore neglected the base-rate information. When the problem story was changed to prevent this bias, subjects' responses shifted from the diagnostic value, yet were still higher than the bayesian answer.

Lyon & Slovic (1976) investigated the effect of various aspects connected with the problem story on the degree of the bias. They investigated aspects of content, extreme base-rate values, presentation order of the information and response format. The base-rate fallacy was observed under all of these conditions.

There are many practical contexts, in which people use diagnostic tools, in order to decide which of two, or more, hypotheses is correct. Such contexts are, for example, medicine, law, and especially intelligence and other military domains. In this cases, ignoring base rates, may have undesirable and severe consequence. In light of the above empirical evidence, it is vital to develop aids in order to direct people in probability assessment of this type.

Fischhoff, Slovic & Lichtenstein (1979) developed the Subjective Sensitivity Analysis (SSA) procedure. Using base-rate problems, previously involved neglect of base-rate, the SSA procedure, directed the subjects to first consider how they would perform the same judgment with various base-rate values, and only then respond. The results showed that this procedure affected subjects's judgment in the sense that they were closer to the normative (Bayesian) answer than is usually found. However, this improvement was not generalized. That is, after solving base-rate problems, using the SSA procedure, subjects had to solve other similarly structured, base-rate problems, without SSA. Again, the base-rates were ignored.

Fischhoff & Bar-Hillel (1984) further investigated the effect of the SSA procedure. They found that although SSA consistently increased usage of base-rate information, it did so as a mechanical procedure, rather than contributed qualitatively to subjects' comprehension.

Fischhoff & Bar-Hillel also tested three alternative techniques of enhancing a variable's salience (focusing techniques): Isolation Analysis (IA), which encourages subjects to consider each information in turn, judging how they would respond if it was the only information available, Minimal Focusing (MF), which instructs subjects explicitly to consider both items of information, and Balanced SSA (BSSA) which applies SSA separately to both items of information. These techniques were effective in changing subjects' performance, in the sense that subjects did not ignore the base-rate information. However, this change can not be attributed to better understanding, since the base-rate information was also considered when subjects responded to other problems not requiring utilizing the base-rate information.

Fischhoff & Bar-Hillel concluded that "It is not enough to motivate subjects or clarify instructions or give problems with a familiar content. In order to improve intuitive judgement a manipulation must constructively change the way in which people concepualize a problem, or give them new cognitive skills with which to examine it. Thus, instead of debiasing procedures, there may be a need for training programs" (p 193).

The effectiveness of various structuring aids, in solving base-rate problems, was investigated by Lichtenstein & MacGregor (1985). Their subjects were required to solve base-rate problems under the following experimental conditions:

a. Control. In this condition, the subjects had to solve the problems without any aid.

b. List. In this condition, the subjects had to list factors, that they believed were relevant to the answer; this was done before answering the problems.

c. Algorithm. In this condition, the subjects were given a full algorithm specifying all the stages of the correct solution. They were required to extract the information from the problem, assign it according to the instructions and do the specified arithmetic.

d. Tutorial. In this condition, the subjects read a seven-page tutorial, specifying the way of solution and explaining why this was the correct one.

The results shown that the list condition had no effect. On the other hand, the algorithm and the tutorial aids did affect performance. However, generalization was observed only for subjects, previously aided by the tutorial. This was manifested by the fact that some of these subjects were able to solve a second base-rate problem, without the aid of the tutorial.

Lichtenstein & MacGregor (1985), concluded that "the tutorial approach holds great promise" (p. 20), since their results had shown that subjects can be taught how to solve base-rate problems successfully, in a relatively short period of time, without individual tutoring, practice, or feedback. This tutorial, however, led to systematic errors in calculating the target probability. In their opinion, this conceptual problem might be rectified by re-writing and expanding the tutorial.

The tutorial aid should be modified, not only in order to over come "built in" errors, but also to increase its effect on the user. In addition, it is important to examine the applicability of this aid to various populations. That is, the subjects who participated in the above were American college students. Different populations have unique and specific characteristics, that may effect the capability of their members to learn and generalize the material, presented to them by such aid. Therefore, it is important to examine this aid, and any modification applied to it, on various populations. One important population is the military one, whose members are potential users of cognitive aids. This population is

characterized by a unique way of thinking, and furthermore, these tasks are usually performed under conditions of stress.

One purpose of the present study is to test the applicability of the tutorial aid, employed by Lichtenstein & MacGregor (1985), to Israeli university students. The second purpose is to modify this tutorial in order to obtain the above goals. An additional purpose is to test the effectiveness of the modified tutorial on Israeli military population, especially under time stress conditions.

The basic concept underlying this modification is Paivio's dual coding hypothesis. According to this hypothesis, learning involve both mental images and verbal processes, operating simultaneously (Paivio, 1971). The tutorial, used by Lichtenstein & MacGregor (1985), verbally explained the base-rate problem and the way of solution. This natural language mediation is a verbal strategy for learning process. Imagery can be used as a non-verbal strategy for learning process. That is, images can be used as a way of organizing verbal items. This method is an effective mode of learning (Adams, 1976). In the modified tutorial, the material is presented both verbally and by images. This is called Training by Mental Image (TbMI). The actual images used, are based on the concept of representation by Ven Diagrams. The application of images can contribute to a better understanding, by turning the somewhat abstract situation, described in base-rate problems, into a more concrete and clear one.

An effective aid is one that improves intuitive performance, and gives new cognitive skills. Such an aid should be effective regardless of variations of content and structure. One content factor, which is another source of emotional stress, that can influence performance, is risk. For example, in solving base-rate problems, the two types of information are interpreted and their relevance is determined. If a problem's content indicate some risk, it may influence these interpretations and the judgement of relevance. This may cause one type of information to be judged as more salient and thus, the availability heuristic may be used. That is, the more salient type of information will be judged as more relevant, and therefore will be the only basis for assessing the required probability. However, If the aid is effective, the salience of one information would not influence the conceptualization of the problem, and the assessed probability.

The effectiveness of the modified tutorial, will also be examined when the problem content indicates various levels of risk. It is hypothesized that if this aid is effective, it will remove the influence of the risk elements.

Experiment III

Experiment III is a partial replication of Lichtenstein & MacGregor, (1985). Of the four experimental conditions employed in the original study, only the algorithm and the tutorial condition were used in the present one. This conditions were selected, since they were found to be effective. The purpose of Experiment III was to test the effectiveness of these aids in solving base-rate problems, on Israeli student population.

Method

Subjects. Sixty students participated in Experiment III. The respondents were recruited from the Tel-Aviv University Introductory Psychology subject pool.

Base-Rate problems. The experiment employed the Light Bulb and Dyslexia problems used by Lichtenstein & MacGregor (1985). The problems are presented in Appendix D. All aspects of the problems were accurately translated into Hebrew.

Experimental design. Two independent variables were manipulated in this experiment: Aid type and Problems' type.

Type of aid. In the first condition, subjects were aided by an algorithm in answering the first base rate problem. This condition is denoted AL. In the second condition, subjects were aided by a tutorial. This condition is denoted TU. Both the algorithm and the tutorial were adopted from Lichtenstein & MacGregor (1985). The algorithm, and the tutorial (Light Bulb version) reported in the original study, were translated to Hebrew. A corresponding algorithm was composed for the Dyslexia version. The algorithm and the tutorial are presented in Appendix E.

Problems' type. As in the original study (Lichtenstein & MacGregor), in one condition, subjects received the Light Bulb problem as a training problem. In the other condition, subjects received the Dyslexia problem as a training problem. Both problems were of similar structure, but the diagnostic

information in the Light Bulb problem related only to the probability of correct diagnosis, while in the Dyslexia problem, it related also to the probability of incorrect diagnosis. For all the subjects the first problem, used as training, was given with either the algorithm or tutorial as an aid. The second problem, used as a generalization problem, was presented without an aid.

The experimental design with two independent variables, which make up 4 deffrent groups, is presented in Table 18.

Table 18: The Experimental Design of Experiment III

TRAINING \ TYPE OF AID	ALGORITHM	TUTORIAL
LIGHT BULB		
DYSLEXIA		

The dependent variables. Three dependent variables were measured in this experiment: response mode, and the degree of confidence and reasonableness for the training problem.

Response mode. Subjects' probability assessments for each problem were classified as follows:

"Correct" - If subject's answer was equal to the normative solution according to Bayes' Theorem.

"Diagnostic" - If subject's answer was equal to the diagnostic value given in the problem.

"Base-Rate" - If subject's answer was equal to the Base-Rate value given in the problem.

"Conditional" - If subject's answer was equal to the Base-Rate value multiplied by the diagnostic value.

"Other" - If subject's answer was not equal to any of the above values.

Degree of confidence. After completing the training problem, the subjects had to rate the degree of confidence they had in the accuracy of their responses, on a scale of 1 to 7, ranging from "not at all confident" (1) to "very confident" (7).

Reasonableness. The respondents answered "yes" or "no", to the question "Does the answer you have reached seem reasonable to you?". If they answered "no", subjects in the algorithm group were also asked to provide a reasonable answer.

Procedure. The subjects were randomly assigned to four groups of 15 students. They were run in groups of 3 to 5 people in small classroom at the university. During the sessions, pocket calculators were supplied, in order to avoid arithmetic errors, which might affect the assessed probabilities, in solving both problems. The experimental sessions were the same for all the groups. Subjects first read the instructions, and then solved the training problem (Light Bulb or Dyslexia) with the aid of the tutorial or the algorithm. All materials used in the performance of the training problem were then collected and subjects were asked to work on the generalization problem (Light Bulb or Dyslexia) without any aid.

Results

Response categories

The distributions of subjects' response mode in both problems whether aided or unaided, are shown in Table 19.

Table 19: Frequencies and Proportions of Subjects' Response Mode for Training and generalization problems

PROBLEM / RESPONSE	TRAINING PROBLEM		GENERALIZATION	
	ALGORITHM	TUTORIAL	ALGORITHM	TUTORIAL
CORRECT	10 35.71%	11 39.28%	16 56.17%	3 10.34%
DIAGNOSTIC	0 0%	1 3.57%	3 10.34%	10 34.48%
BASE RATE	9 32.14%	9 32.14%	3 10.34%	3 10.34%
CONDITIONAL	0 0%	0 0%	1 3.44%	6 20.68%
OTHER	9 32.14%	7 25%	6 20.68%	7 24.13%

The data in Table 19 show that the proportion of "correct" responses for the training problem is almost the same for both Al (39%) and TU (36%) groups. On the other hand, when considering the generalization problem, this proportion is much higher for the TU group (56%) than for the AL group (10%). For both aid types, the proportion of "correct" responses is higher for the Light Bulb problem than for the dyslexia problem, when given as training problems. When given as generalization problems, this proportion is equal for the group aided by the algorithm. For the tutorial group this proportion is much higher for the Light Bulb problem than for the Dyslexia one.

A chi-square test performed on the response frequencies for the training problem, failed to reach significance. A significant effect was obtained, however, for the generalization problem (chi-square=16.31, df=4, p<.01).

Confidence

The mean confidence ratings are shown in Table 20, for all subjects in each aiding condition and for each base-rate problem (across groups).

Table 20: Means of Confidence Ratings for Training Problems

GROUP / PROBLEM	TUTORIAL	ALGORITHEM	ACROSS GROUPS
LIGHT BULB	3.33	5.27	4.3
DYSKEXIA	3.73	4	3.86
ACROSS PROBLEMS	4.66	3.35	4.8

Table 20 indicates that subjects reported higher confidence in the accuracy of their answers to the Light Bulb problem than of the answer to the Dyslexia problem. Subjects in the TU groups reported higher confidence than those in the AL groups.

A t-test performed on the difference between confidence ratings of the two groups was found to be significant (t=2.45 df=57 p<.01). The difference between confidence ratings for Light Bulb and Dyslexia problem failed to reach significance.

Reasonableness

The proportions of "yes" and "no" answers to the question "Does the answer you have reached seem reasonable to you?" are shown in Table 21.

Table 21: Proportions of Reasonable and Unreasonable Responses for Training problems

AID / RESPONSE	LIGHT BULB			DYSLEXIA			BOTH PRBLEMS		
	ALGORITHM	TUTORIAL	TOTAL	ALGORITHM	TUTORIAL	TOTAL	ALGORITHM	TUTORIAL	TOTAL
NO	7 46.66%	2 14.28%	9 31%	5 35.71%	1 8.33	6 23%	12 41.37%	3 11.53%	15 27%
YES	8 53.33%	12 85.71%	20 69%	9 64.28%	11 91.66%	20 78%	17 58.63%	23 88.46%	40 73%

The proportion of subjects reporting of reasonable answers (answered "yes" to the above question) was higher for the Tutorial group (41%) than the Algorithm group (12%). A chi-square test on these data was significant (chi-square=4.74, df=1, $p<.05$).

Discussion

Experiment III examined the effectiveness of the algorithm and tutorial aid in solving base-rate problems, on an Israeli students sample. The results showed that both the algorithmic aid and the tutorial aid were effective, as direct aid, and led to higher proportion of correct responses. However, the tutorial aid led to essential change in thinking and in the way subjects' conceptualized the problems, as manifested by the generalization observed under this condition. The poor generalization found for the algorithmic aid, indicates that this aid was technical and did not give the subjects new cognitive skills with which to examine the base-rate problem.

This is also supported by the degree of confidence and reasonableness, reported under the two aiding conditions. Subjects aided by the tutorial reported higher degree of confidence than those aided by the algorithm. The proportion of subjects reporting of reasonable answers was higher for the subjects aided by the tutorial than for those aided by the algorithm.

algorithm.

An interaction existed between the aiding method and problem type which determined training effectiveness. The tutorial aid was more effective, in solving the Light Bulb problem, than in solving the Dyslexia one. Had this difference between the two generalization problems been observed under both aiding conditions, it would have indicated that this was a result of the different content and structure of the problems. Since this is not the case, it may indicate that the explanation was not clear enough and had only limited contribution to the understanding of the situation described in base-rate problems.

Experiment IV

Experiment III examined the effectiveness of the algorithm and tutorial for aiding Israeli students in solving base-rate problems. The results showed that, although both aids were effective, only the tutorial led to generalization of the correct way of solution. Experiment IV was designed to further develop this aid by introducing mental images. This was tested on military population, using base-rate problems of military content, and under normal and time stress conditions.

Method

Subjects. Two hundreds twenty-two IDF maintenance junior officers participated in Experiment IV. The subjects have had secondary education.

Base-Rate Problems. All the subjects were asked to solve 6 different Base-Rate problems. The problems are presented in Appendix F. All problems were similarly structured. Each contained base-rate information about a certain phenomenon, information about the degree of accuracy of a given diagnostic device, and a question concerning the accuracy of a particular diagnosis. Subjects were to answer in terms of a probability or a percentage as discussed below.

The Color Blindness problem was used for illustration, the Parachute and Jaundice problems were used as training, and the Missile, Seals and Masks problems were generalization problems.

The three generalization problems were of a military content, which was relevant to the subject pool, and were presented in three different forms specifying different levels of risk: neutral, general and personal (see Appendix F). The

Missile problem read as follows:

Intelligence sources revealed that a hostile Army had purchased sophisticated G-7 anti aircraft missiles.

Israeli industry had developed a special device, capable of receiving the signals broadcasted from anti aircraft missiles that enabled identification of missile type. The device is known to be accurate in 80% of the cases, that is, a G-7 missile type and missiles of "other types" will be correctly identified as such in 80% of the cases. Knowledge of the exact type of missile, improves the defense profile an aircraft flying in a bound area.

Researches done by the Tactical Warfare Development Committee show that the chances for launching a G-7 type missile is 10%.

"An anti aircraft missile has been sent to a certain are and was identified by the device as being of type G-7".

<u>What are the chances that this missile is really a type G-7 missile</u>?

<u>Experimental design</u>. There were three independent variables: Provision of an aid, time restriction and levels of risk.

The TbMI Aid: In the aided condition, (denoted A) the subjects were presented with a tutorial, which was a modified version of the one used by Lichtenstein & MacGregor (1985). The modification involved changing only the mode of presentation of the problem situation, <u>not</u> the method of calculation (which was an expansion of the explanation of base-rate problems given by Beyth-Marom, Dekel, Gombo, and Shaked, 1985). The Tutorial (shown in Appendix G) contained a detailed analysis of the situation described in the Color Blindness problem, represented pictorially. A "tree" structure was used as the method of focusing the subject on the target sub-populations. (i.e., those who were diagnosed as color blind, and of those diagnosed as such, those who were, in fact, color blind). The tutorial was accompanied by verbal explanation and slides. This was followed by the two training problems (Parachute and Jaundice). After completing the first training problem, these subjects were given feedback by showing them the correct solution. In the unaided condition, (denoted UA) the subjects read a short essay discussing general statistical subjects.

Risk: In the neutral condition (denoted NR), the subjects had to solve generalization problems describing situations of neutral risk. By "Neutral risk" is meant that in the problem it was not indicated of any danger to the reader himself or to relevant others. For example:

> "An anti aircraft missile has been sent to a certain are and was identified by the device as being of type G-7".

In the general risk condition (denoted GR), the subjects had to solve generalization problems specifying general risk. By "General Risk" is meant that in the problem a hint of potential danger to some one other than the reader him/herself was given. For example:

> "An anti aircraft missile has been launched to a certain area where only Israeli aircraft fly, and was identified by the device as being of type G-7".

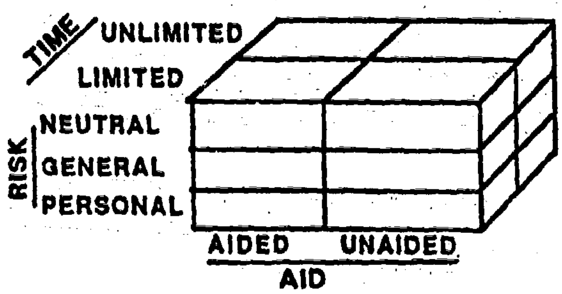
In the personal risk condition (denoted PR), the subjects were given generalization problems specifying personal risk. By "Personal Risk" is meant that the problem described situations endangering the reader himself. For example:

> "Suppose you are a pilot flying an Israeli aircraft. An anti aircraft missile that had been launched to the area where you are flying was identified by the device as being of type G-7".

Time restriction: There were two time restriction conditions. In the unlimited time condition (denoted NTL), the subjects were to solve each generalization problem without any time restriction. In the limited time condition (denoted TL), the subjects had to solve each generalization problem within 4 minutes. The 4 minute limit was determined by running the NTL/A groups first, and measuring the time required to solve each one of the generalization problems. The time limit for the "time restricted" groups was chosen by taking the lowest time-to-solution with the highest frequency, provided that two or more subjects achieved that time-to-solution.

The experimental design with three independent variables, which make up 12 groups, is presented in Table 22.

Table 22: <u>The Experimental Design of Experiment IV</u>



<u>The dependent variables</u>. Four dependent variables were measured in this experiment: response mode, the degree of confidence in the accuracy of their answer, reasonableness and subjective mental load.

Response mode. The answers to each generalization problem were classified in the same manner as done in Experiment III.

Confidence. The degree of confidence in the accuracy of each response was rated in the same manner as in Experiment III.

Reasonableness. As in Experiment III, the respondents answered "yes" or "no", to the question "Does the answer you have reached seem reasonable to you?".

Subjective mental load. After completing all three generalization problems, subjects filled the subjective mental load questionnaire, used in Experiments I and II.

<u>Procedure</u>. The subjects were randomly assigned to 12 groups of 15 to 20 officers, each was run separately in small class rooms at an army base. The instructions preceding each session indicated that the purpose of the study was to examine the ways in which people solve various problems. It was emphasized that participation in the experiment was anonymous, and that subjects' performance would not affect their career. During the sessions, pocket calculators were supplied for use, in order to avoid arithmetic errors in solving the problems. The experimental sessions were the same for all the groups. Subjects were given a brief introduction to the study and then filled in a standard form containing details such as age, sex, months of service, current job, command experience, education and the like.

All subjects were first administered the Color Blindness problem. After solving this problem the subjects read the tutorial or the general essay, accompanied by the verbal presentation and the slides. This was followed by the training problems. During training, the subjects who read the tutorial were provided with feedback, and were allowed to ask questions. All material used in the performance of the above problems were then collected, and subjects were asked to work on the generalization problems, again unaided. At this point, the risk and time restriction conditions were manipulated, without prior notice, i.e., subjects were not told at the beginning of the experimental session, that they would be time-restricted later. The respondents final task was to fill in the subjective mental load questionnaire.

Results

Validity of Time Stress manipulation

The time stress manipulation was validated based on subjects' subjective ratings of time stress (Appendix B, item 5). The mean ratings and s. d. for all the groups are presented in Table 23.

Table 23: Means and standard Deviations of Time Stress Rating (Upper No.= Mean; Lower No.= s.d.

AID / TIME	AIDED	UNAIDED	TOTAL
UNLIMITED	1.81	1.45	1.61
	1.53	1.14	1.34
LIMITED	2.78	1.58	2.17
	1.62	1.10	1.50
TOTAL	2.31	1.51	1.89
	1.64	1.12	1.44

Table 23 shows that the mean rating for all the TL groups across the "aid" and "risk" conditions is higher than the mean rating for the NTL groups.

An analysis of variance performed on these data showed a significant main effect for time limit (F(1,192)=7.88, p<.01). The aid x time limit interaction was also found to be significant (F(1,192)=5.06), p<.01). This interaction is shown in Figure 1.

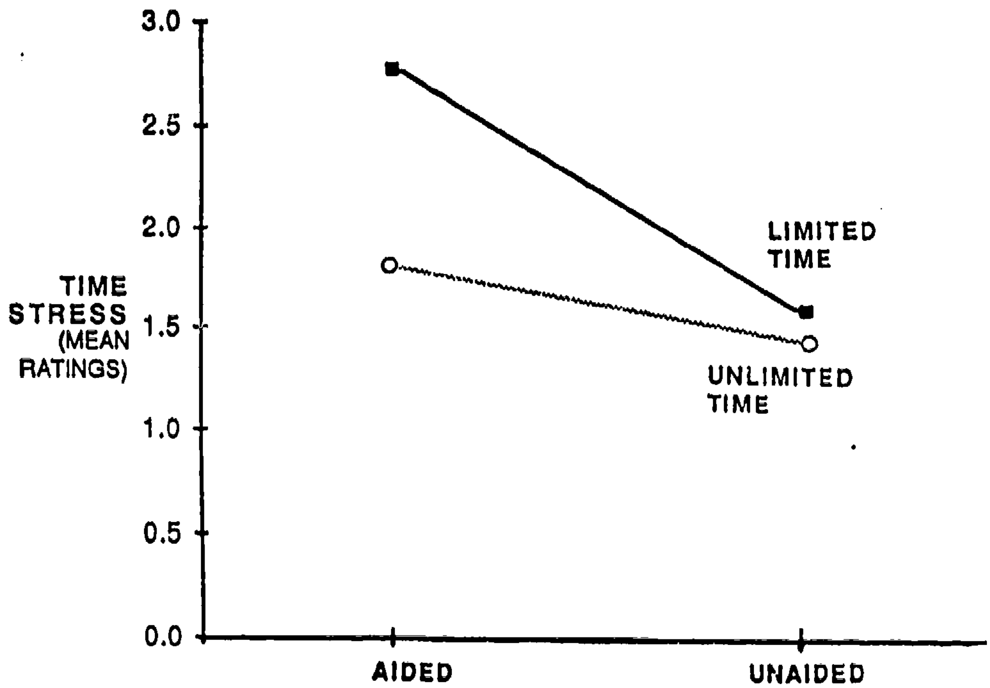


Figure 1: Mean Time Stress Ratings as Function of Time Restriction and Aiding Conditions

Response Categories

The Aiding Manipulation

A and UA groups. The distribution of subjects' response mode for each generalization problem is shown in Table 24.

Table 24 shows that in the UA groups, the most frequent response category is the "diagnostic" one (56% for Missile, 53% for Seals and 49% for Masks), while the proportion of "correct" responses is always 0. In the A groups the most frequent category is the "correct" one (48% for Missile, 62% for Seals,

and 65% for Masks), while the proportion of "diagnostic" responses is very low. Note that the proportion of "conditional" responses in the A groups (21% for Missile, 20% for Seals and 23% for Masks) decreased relative to the UA groups.

Table 24: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	53 48.18%	0 0%	67 62.03%	0 0%	66 65.34%	0 0%
DIAGNOSTIC	3 2.72%	61 69.31%	5 4.62%	59 53.15%	7 6.93%	54 49.09%
BASE RATE	0 0%	5 5.68%	2 1.85%	3 2.7%	4 3.96%	5 4.54%
CONDITIONAL	4 3.36%	23 26.13%	3 2.77%	22 19.81%	1 0.99%	25 22.72%
OTHER	50 45.45%	20 22.72%	31 28.7%	27 24.32%	23 22.77%	26 23.63%

A chi-square test on these data indicated a significant difference between the response distributions of the A groups and the UA groups for Missile (chi-square=136.78, df=4, p<.01), Seals, (chi-square=127.46, df=4, p<.01), and Masks (chi-square=124.50, df=4, p<.01).

The proportion of correct responses across all three generalization problems was computed for each subject. This proportion is, on the average .57 for all the A groups and 0 for all the UA groups. Analysis of variance on these data showed a significant main effect for aid type (F(1,162)=248.858 p<.01).

NTL/A and NTL/UA groups. The distribution of subjects' responses, with regard to the NTL groups, to each generalization problem is shown in Table 25.

The data in Table 25 show that the proportion of "correct" responses is higher for the A/NTL groups than for the UA/NTL groups, and the proportions of "conditional" responses in the A/NTL groups decreases relative to the UA/NTL groups.

Table 25: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for NTL Groups

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	33 58.92%	0 0%	37 67.27%	0 0%	35 67.3%	0 0%
DIAGNOSTIC	3 5.35%	32 57.14%	3 5.45%	28 50%	4 7.69%	23 41.07%
BASE RATE	0 0%	4 7.14%	0 0%	3 5.35%	1 1.92%	2 3.57%
CONDITIONAL	1 1.78%	13 23.21%	2 3.63%	14 25%	0 0%	19 33.92%
OTHER	19 33.92%	7 12.5%	13 23.63%	11 19.64%	12 23.07%	12 21.42%

A chi-square test performed on these data indicated a significant difference between response distributions of the A NTL groups and the UA NTL groups for Missile (chi-square=76.85, df=4, p<.01), Seals (chi-square=69.32, df=4, p<.01) and Masks (chi-square=67.64, df=4, p<.01).

The distribution of subjects' responses, when considering the TL groups, to each generalization problem is shown in Table 26.

Table 26: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for TL Groups

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	20 37.03%	0 0%	30 56.6%	0 0%	31 63.26%	0 0%
DIAGNOSTIC	0 0%	29 54.71%	2 3.77%	31 56.36%	3 6.25%	31 57.4%
BASE RATE	0 0%	1 1.88%	2 3.77%	0 0%	3 5.25%	3 5.55%
CONDITIONAL	3 5.55%	10 18.86%	1 18.86%	8 14.54%	1 2.08%	6 11.11%
OTHER	31 57.4%	13 24.52%	18 33.96%	16 29.09%	11 22.91%	14 25.92%

Again, the data in Table 26 indicate that the proportion of "correct" responses is higher for the A/TL groups than for the UA/TL groups, and the proportions of "conditional" responses in the A/TL groups decreases relative to the UA/TL groups.

A chi-square test on these data indicated a significant difference between response distributions of the A groups and the UA groups for Missile (chi-square=61.13, df=4, p<.01), Seals (chi-square=63.03, df=4, p<.01) and Masks (chi-square=57.88, df=4, p<.01).

Risk. The distribution of subjects' response mode, in the various risk levels groups, to each generalization problem is shown in Tables 27 to 29.

Table 27: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for Neutral Risk Groups

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	21 58.33%	0 0%	21 58.33%	0 0%	22 70.96%	0 0%
DIAGNOSTIC	0 0%	0 0%	2 5.55%	22 61.11%	1 3.22%	21 63.63%
BASE RATE	1 2.77%	22 61.11%	0 0%	1 2.77%	0 0%	6 18.18%
CONDITIONAL	1 2.77%	5 13.88%	0 0%	4 11.11%	0 0%	0 0%
OTHER	13 36.11%	9 25%	13 36.11%	9 25%	8 25.8%	6 18.18%

Table 28: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for General Risk Groups

RESPONSE \ PROBLEM	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	16 45.71%	0 0%	23 67.64%	0 0%	19 57.57%	0 0%
DIAGNOSTIC	1 2.85%	19 51.35%	1 2.94%	16 43.24%	4 12.12%	16 59.25%
BASE RATE	0 0%	2 5.4%	0 0%	1 2.7%	1 3.03%	1 3.7%
CONDITIONAL	2 5.71%	9 24.32%	1 2.94%	10 27.02%	1 3.03%	9 33.33%
OTHER	16 47.71%	7 18.91%	9 26.47%	10 27.02%	8 24.24%	1 3.7

Table 29: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for Personal Risk Groups

RESPONSE \ PROBLEM	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
CORRECT	16 41.02%	0 0%	23 60.52%	0 0%	25 67.56%	0 0%
DIAGNOSTIC	1 2.56%	20 55.55%	2 5.26%	21 55.26%	2 5.4%	17 44.73%
BASE RATE	0 0%	3 8.33%	2 5.26%	1 2.63%	3 8.1%	2 5.26%
CONDITIONAL	1 2.56%	9 25%	2 5.26%	8 21.05%	0 0%	10 26.31
OTHER	21 53.84%	4 11.11%	9 23.68%	8 21.05%	7 18.42%	9 23.68%

As for Time limit and Aid, the data in Tables 27 to 29 indicate that the proportion of "correct" responses is higher for the A groups than for the UA groups, and the proportions of "conditional" responses in the UA groups decreased, for all risk levels.

A chi-square test on the data for each risk level, indicated a significant differences between response distributions of the A groups and the UA groups for each generalization problem. The test results are summarized in Table 30.

Table 30: Chi-Square Tests Results of Subjects' Response Mode, in All Generalization Problems, for Each Risk Level (Upper No.=Chi-Square, Middle No.=df Lower No.=Significance)

PROBLEM \ RISK	NEUTRAL	GENERAL	PERSONAL
MISSILE	43.57 3 0.00	42.15 4 0.00	54..12 4 0.00
SEALS	43.39 4 0.00	44.66 4 0.00	42.68 4 0.00
MASKS	48.20 4 0.00	32.95 4 0.00	47.29 4 0.00

The Time Limit Manipulation

This manipulation seemed to have only a minor effect on response distributions. The response distribution for the TL groups and the NTL groups are shown in Table 31. A chi-square test comparing the responses distribution of the TL time and NTL groups failed to reach significance, except for Missile problem (chi-square=10.10, df= 4, p<.05). This indicates a greater proportion of "correct" and "diagnostic" responses in the NTL group, and a higher proportion of "other" responses in the TL groups. Significance was reached for the following specific comparisons:

Table 31: Frequencies and Proportions of Subjects' Response Mode for Each Generalization Problem for TL and NTL Groups

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	LIMITED	UNLIMITED	LIMITED	UNLIMITED	LIMITED	UNLIMITED
CORRECT	33 29.5%	20 18.7%	37 33.3%	30 27.8%	35 32.4%	31 30.1%
DIAGNOSTIC	35 31.3%	29 27.1%	31 27.9%	33 30.6%	27 25.0%	34 33.0%
BASE RATE	4 3.6%	1 0.9%	3 2.7%	2 1.9%	3 2.8%	6 5.8%
CONDITIONAL	14 12.5%	13 12.1%	16 14.4%	9 8.3%	19 17.6%	7 6.8%
OTHER	26 23.2%	44 41.1%	24 21.6%	34 31.5%	24 22.2%	25 24.3%

For the Missile problem only - the proportion of "correct" responses in the A groups, is higher for the NTL groups than the TL groups (chi-square=10.04, df=3, $p<.05$).

For the Seal problem only - in the A/NR groups, the proportion of "correct" responses is higher for the NTL groups (chi-square=11.57, df=4, $p<.05$).

For the Masks problem only - in the UA groups, the proportion of "diagnostic" responses is lower for the NTL groups, while the proportion of "conditional" responses is much higher (chi-square=8.27, df=3, $p<.05$). In addition, in the UA/GR groups, the proportion of "diagnostic" responses is lower for the NTL groups, while the proportion of "conditional" responses is higher (chi-square=8.76, df=3, $p<.05$).

The Risk Manipulation

The risk manipulation was found to have no effect on response distribution under any of the conditions.

Confidence

The mean confidence ratings and s. d. for each generalization problem are shown in Table 32.

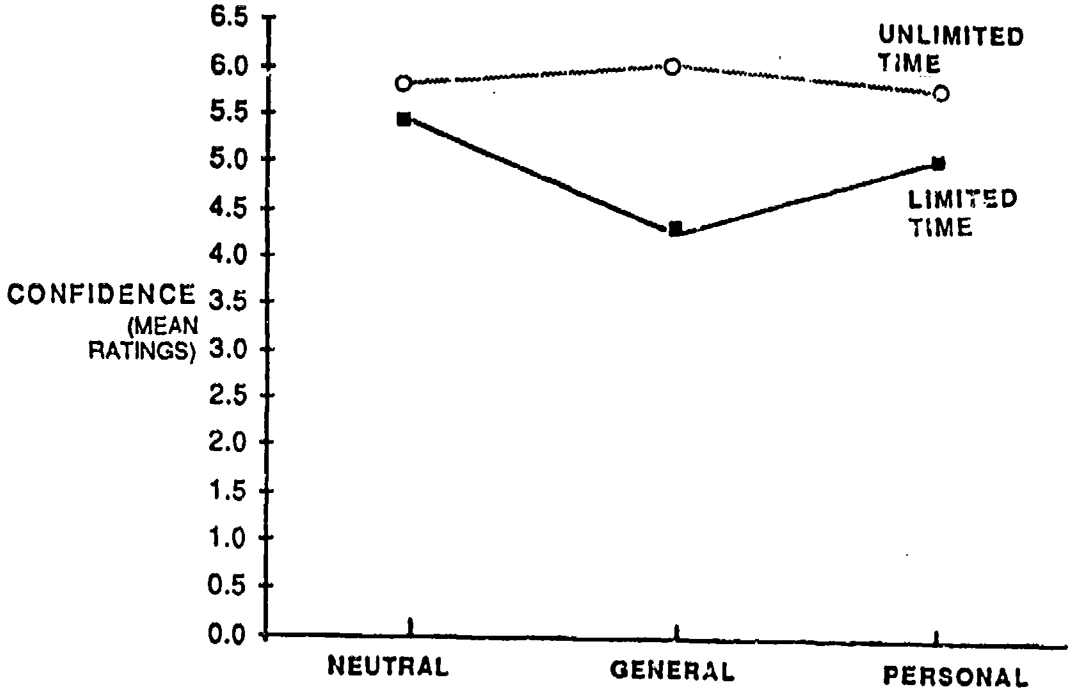
Table 32: Means and s.d. of Confidence Ratings For all Generalization Problems (Upper No.=Mean; Lower NO.= s.d.)

PROBLEM / RESPONSE	MISSILE		SEALS		MASKS	
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED
LIMITED	5.42 1.65	5.02 1.67	5.67 .89	5.02 1.56	5.76 1.00	5.10 1.34
UNLIMITED	5.17 1.51	5.58 1.67	6.17 1.33	5.69 1.39	6.29 1.64	5.86 1.49

Table 32 shows that, for all three generalization problems, the mean ratings are higher for the NTL groups than the TL groups. The mean ratings are higher for the A groups than for the UA groups, for the missile and masks problems. For the seals problem, the mean ratings are higher for the UA group then the A group under time stress conditions. With regard to the risk groups, the ratings are highest for the NR groups, followed by the GR and then the PR groups.

Analyses of variance on these data for each generalization problem, indicated significant main effect of time limit, for the Seals ($F(1,176)=8.71$, $p<.01$) and Masks ($F(1,176)=10.51$, $p<.01$) problems. A significant main effect of aid condition was also found for the Seals ($F(1,176)=8.45$, $p<.01$) and Masks ($F(1,176)=7.58$, $p<.01$) problems. The two-way interaction effect of time limit x risk was significant for the Seals problem ($F(1,176)=5.23$, $p<.01$). This interactions is shown in Figure 2.

Figure 2: Mean Confidence Ratings as Function of Time Restriction and Risk Conditions



An overall measure of confidence was obtained by averaging the ratings across all three generalization problems. The mean confidence ratings and s. d. for the various groups are shown in Table 33.

Table 33: Means and s.d. of Confidence Ratings for all Experimental groups (Upper No.=Mean; Lower NO.= s.d.)

PROBLEM / RESPONSE	NEURTAL		GENERAL		PERSONAL		TOTAL
	AIDED	UNAIDED	AIDED	UNAIDED	AIDED	UNAIDED	
LIMITED	5.87 1.04	5.42 1.22	5.76 1.19	4.89 1.62	5.13 1.03	5.30 .80	5.38 1.28
UNLIMITED	5.94 1.45	5.58 1.28	6.28 1.04	5.98 1.06	5.93 .98	5.82 1.37	5.86 1.22
TOTAL	5.90 1.23	5.52 1.28	5.92 1.18	5.43 1.56	5.45 1.17	5.54 1.11	5.60 1.26

APPENDIX A

Estimation problems

V1. How much food (in Kg.) does an average percon consume during his entire lifetime.

V2. What is the number of beds in all the general hospitals in Israel?

V3. How many liters of water, for home usage, are consumed in Israel during one year?

V4. How many cars are onwed by the Israeli population?

V5. How many students graduate scondary school in a year?

V6. How many airplanes land and takeoff (in season) in one day at the Ben-Gurion Airport?

V7. What is the number of members of the academic stuff employed in Israeli universities?

V8. How many active bus drivers are employed in "EGED"?

APPENDIX B

Subjective mental load questionnaire

1. What was the degree of difficulty you encountered in answering the questions?

Very Easy |___|___|___|___|___|___| very difficult

2. How much thinking effort was required to answer the quesitons?

Little Effort |___|___|___|___|___|___| A Lot of Effort

3. Was answering the questions tiring?

Not Tiring at all |___|___|___|___|___|___| Very Tiring

4. Was answering the question frustrating

Not Frustrating at all |___|___|___|___|___|___| Very Frustrating

5. Did you have enough time for answering the questions?

Enough Time |___|___|___|___|___|___| Not Enough Time

APPENDIX C

Training for Using Algorithmic Decomposition

In many situations in everyday life we have to make a vast amount of decisions. Each decision is usually based on a set of data concerning different aspects of the specific situation at hand. For example, a commander has to decide how to deploy his forces in a certain area. To make such a decision, he needs a large amount of data. For instance, he has to know the number of soldiers in his force; how many arms are at his disposal; the amount of available ammunition; what are the terrain conditions; what is the enemy's troop deployment; etc.

That is to say that any and every decision must be made after taking into account the answers to a series of questions, some of them quantitative. Some of these answers are readily available and easily obtained, e.g. the number of arms and soldiers in the commander's force. Other relevant information can be obtained through various military services. But in many cases the necessary data is unavailable. In these cases we must estimate the values.

If reliable decisions are to be made, then the estimates must be made as accurately as possible. In addition, in many decision making situations the time factor is very crucial, and thus the estimates must also be made as quickly as possible. For example, if the commander errs in underestimating the enemy's arms, he may decide on a force deployment that endangers his soldiers. Similarly, if he spends too much time in gathering the relevant information his decision, although correct, may be made too late.

Therefore, it is important that he use a method which will help him reach the most precise estimates possible, in the shortest possible time. One of the possible methods is the use of partial knowledge to estimate related quantities, and applying this to generate the target estimate.

Research has shown that people tend to make lot of mistakes when required to estimate unknown quantities. To be more accurate and to avoid errors one has to use an efficient method of estimation.

In this experiment you will be presented with a method based on utilizing partial knowledge or sub-estimates of related quantities. The method involves the following three elements:

1. The sub-division of the target question into a number of sub-questions.

2. Assigning values to the sub-questions.

3. Combining these values, by rule, to arrive at the target answer.

The method is illustrated using the following example. Suppose the question is: What is the number of beds in a certain general hospital?

Since it is unlikely for this answer to be known, the correct answer must be estimated. To do this the target question must first be divided into sub-questions that are related to the target answer. For example:

1. How many departments are there in a general hospital?

2. How many rooms are there in each department?

3. How many beds are there in each room?

In the next stage we can try to answer these questions. The appropriate values may be available or more easily estimated than the target value. For instance, based on the knowledge we have, we can count the names of various department and reach an accurate estimate. In the same manner we can estimate the number of rooms in each department and the number of beds in each room.

The next stage is to define a rule for combining the various values we have reached in order to arrive at the target estimate. In our example, the rule is:

1. Multiply the number of beds in each room by the number of rooms in each department. The result is the number of beds in each department.

2. Multiply the number of beds in each department by the number of departments in the hospital.

In this way we obtain a value which is an accurate estimate of the target value. This value may be not identical to the

target value, but it is probably a good approximation of the real value and more accurate than any guess.

The set of sub-questions and the rules for combining the estimated values are called Algorithm.

There are many ways to compose an algorithm. For example one can retrieve from one's memory partial knowledge which is relevant to the target value, and define a rule for combining these pieces of knowledge. It is also possible to develop a set of sub-questions and a rule, and estimate the values required by the sub-questions.

Algorithms can involve sub-estimates of values that are larger or smaller, in magnitude, than the target value. Algorithms can involve integers or fractions (proportions), e.g., the proportion vs. the number of smokers in Israel.

In defining an algorithm one has to use appropriate measurement units, for example the distance between point A and point B is best estimated in Km. than in Cm.

It is advised to avoid composing very long algorithms, since a long one may increase the error in the target value, and lengthen the time required to reach it.

The above method can be applied when estimating the following quantity:

<u>How many cigarettes are manufactured in Israel in a year?</u>

We will define an algorithm that will aid us in making as accurate an estimate as possible. Work according to the following stages:

a. <u>Locate relevant knowledge domain</u>.

At this stage we have to locate and count relevant domains of knowledge, on which we can base rules of calculations in order to answer the target questions. We therefore have to find topics for which we have available knowledge. For example, we can define rules of calculations according to one of the following domains of knowledge:

1. The consumption of cigarettes in Israel.

2. The production capacity of cigarette manufacturers in Israel.

3. The amount of cigarettes sold in Israel.

b. <u>Determine the knowledge domain on which the algorithm will be based</u>.

At this stage we have to examine each domain of knowledge, and decide which one offers the most available information. For example, we will examine the amount of information required for estimating how many cigarettes are manufactured in Israel in a year, based on to the consumption of cigarettes in Israel. This information can be as follows:

1. The number of smokers in Israel.

2. The number of cigarettes consumed by each smoker in a certain period of time.

Now we will examine the amount of information required for this estimation, based on to the production capacity of cigarette manufacturing factories in Israel. This information can be as follows:

1. The number of cigarette manufacturers in Israel.

2. The production of each factory in a certain period of time.

Finally, we will examine the amount of information required for this estimation, based on the amount of cigarettes sold in Israel. This information can be as follows:

1. The number of stores selling cigarettes.

2. The amount of cigarettes sold in each store during a certain period of time.

Obviously the information available to each one of us is different, and it is possible that one may prefer to compose algorithms based on a certain knowledge domain, while the other may prefer another domain. It is likely that, for most of us, the most available and accurate information of all three knowledge domains, discussed above, is the one related to the consumption of cigarettes in Israel. It would seems easier to estimate the number of cigarettes consumed by each smoker,

and thus the consumption of all smokers, than to count all the stores and the amount of cigarettes sold in each one.

Therefore we will choose knowledge domain (1): The consumption of cigarettes in Israel.

c. Locate a basic information unit.

One we choose the knowledge domain on which the algorithm will be based, we can start composing it. First we must locate a basic information unit that will serve as the starting point. This unit must follow the following criteria:

1. It must be relevant to the target question.

2. It can be assigned an estimable value.

3. Its value can be changed by adding new information.

A basic unit can be, for example, the average amount of cigarettes consumed by one smoker per day. This will determine one of the sub-questions in the algorithm.

d. Compose the algorithm

Based on the sub-question defined above, we will now define other sub questions, each referring to any information that may bring the initial value closer to the target estimate. Also, we will define the rules of calculations according to which the sub-estimates are combined.

The resulting algorithm may be as follows:

How many cigarettes are manufactured in Israel in a year?

a. What is the population of Israel?

b. What proportion of the population smokes?

c. What is the number of smokers in Israel?
[Multiply (a) by (b)].

d. How many cigarettes does the average smoker consume per day?

e. How many cigarettes are consumed in Israel per day?

[Multiply (c) by (d)].

f. How many days are there in a year?

g. How many cigarettes are consumed in Israel in a year?
[Multiply (e) by (f)].

e. Making the estimation

Now we will try to make the sub-estimates and generate the target estimate. Please work according to the above algorithm.

After you have completed the cigarettes question, consider the following question:

How many Kg. of fish are caught in Israel in one year?

We will present you two different algorithms for making this estimation. Read through the algorithms carefully and decide which one is more effective and yield a more accurate estimate.

Algorithm 1.

a. How many fishermen are there in Israel?

b. How many Kg. of fish are caught by one fisherman in one day?

c. How many Kg. of fish are caught by all the fishermen in Israel in one day?
[Multiply (a) by (b)]

d. How many working days are there in one year?

e. How many Kg. of fish are caught in Israel in one year?
[Multiply (c) by (d)].

Algorithm 2.

a. What is the population of Israel?

b. How many Kg. of fish are consumed by one person in one year?

c. How many Kg. of fish are consumed by the entire population in one year?
[Multiply (a) by (b)]

d. What is the proportion of imported fish of all fish consumed by the entire population in one year?

e. How many Kg. of fish are imported each year? [Multiply (c) by (d)]

f. How many Kg. of fish are caught in Israel in one year? [Subtract (e) from (c)].

Which of the previous two algorithms seems more effective, and why? ____
____
____
____
____

When you have finished please report to the experimenter.

Explanation: From among the two algorithms presented above, the more effective would seem to be Algorithm 2. The answers to the sub-questions of Algorithm 2, are more available than those to Algorithm 1.

In the course of this experiment you will be presented with a number of estimation problems similar to those you have solved. You are required to solve these problems according to the method described above.

Before proceeding please reread the method, and if you have any questions ask the experimenter.

APPENDIX D

Base-rate problems for Experiment III

The Light Bulb Problem

A light bulb factory uses a scanning device which is supposed to put a mark on each defective bulb it spots in the assembly line. Eighty-five percent of the light bulbs on the line are OK; the remaining 15% are defective.

The scanning device is known to be accurate in 80% of the decisions, regardless of whether the bulb is actually OK or actually defective. That is, when a bulb is good, the scanner correctly identifies it as good 80% of the time. When a bulb is defective, the scanner correctly marks it as defective 80% of the time.

suppose someone selects one of the light bulbs from the line at random and gives it to the scanner. The scanner marks this bulb as defective.

What is the probability that this bulb is really defective?

The Dyslexia Problem

Dyslexia is a disorder characterized by an impaired ability to read. Two percent of all first graders have dyslexia. A screening test for dyslexia has recently been devised that can be used with first graders. The screening test is cheap and easy to administer; it identifies those children who will later be given a more extensive test to determine for sure whether the child has dyslexia. The screening test is not completely accurate. For children who really have dyslexia, the screening test is positive (indicating dyslexia) 95% of the time. But it also gives a positive (dyslexia) result for 5% of the normal children, the ones who do not have dyslexia.

A first grader is given the screening test and the result is positive, indicating dyslexia.

What is the probability that the child really has Dyslexia?

APPENDIX E

Tutorial (Light bulb version)

Consider the following problem:

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

a. 90% of the cabs in the city are Green and 10% are Blue.

b. A witness identified the cab as Blue.

The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 70% of the time and failed 30% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

Research has shown that people often have trouble answering problems like this. In this portion of today's experiment, we are presenting you with a mini-tutorial to see if instruction will help you solve such problems. Please read through the tutorial carefully. We have allowed time in the experiment for you to do that.

Tutorial

The class of problems here addressed are problems for which two kinds of information are given and a probability is requested. One kind of information is about the population or populations in question. The other kind of information is specific to the case at hand.

In the problem given above, the population is the population of cabs in the city. The population information is that 90% of the cabs are Green and 10% are Blue. The specific information concerns the specific cab that was involved in a hit and run accident. The witness said that the specific cab was Blue. But we also know about this testimony that the witness is not perfectly accurate. The witness is able to correctly identify the color of the cab 70% of the time.

The way most people usually go wrong in solving these

problems is that they concentrate too much on the <u>specific</u> information and tend to neglect the <u>population</u> information. Maybe the specific information seems more immediately relevant to them. Or perhaps they just don't know how to go about combining the information to produce a single answer. Here is a way of doing just that:

<u>Step 1. Draw a table</u>. Begin by drawing a "two-by-two" table, that is, a diagram with two rows and two columns, like this:

<u>Step 2. Label the table</u>. We'll label the columns for the population information. The population is cabs in the city, which are either Blue or Green. The rows get the specific information, that is, the witness testimony, which was Blue--but for completeness, we'll also label the other row Green, because the witness <u>could</u> have said Green. No now our table looks like this:

		Cabs in the City	
		Blue	Green
Witness said:	Blue		
	Green		

Labeling the table is not quite as simple as it may first appear. Notice that the sub-labels, "Blue" and Green", are the <u>same</u> for the rows and the columns. This should generally be true in such problems. It would be a mistake to label the rows according to whether the witness was accurate or inaccurate:

Witness said: Right

Wrong

The problem <u>could</u> be solved with such labeling, but <u>not</u>

using the method we are teaching you here. In general, the sub-labels are the two possible states of the world. The main labels (e.g., "Cabs in the City" and "Witness said:") indicate the source of information. One source is always population information (here, the relative number of cabs in the city); the other source is always specific information (here, what the witness said).

Notice that if there were numbers in the four cells of the table, we could calculate row totals and column totals and a grand total for the whole table. The places for these totals are shown below with dashed lines.

Cabs in the City

		Blue	Green	Row Totals:
Witness said:	Blue			------
	Green			------
Column Totals:		-----	-----	------ Grand Total

Step 3. Assign an arbitrary grand total. To get started, we'll fill in the grand total. That should be the total number of cabs in the city. But we don't know how many cabs there are in the city. So we pick an arbitrary total of 1,000. We could use 10 or 100 (or any other number), but using 1,000 will make later calculations easier.

Cabs in the City

		Blue	Green	
Witness said:	Blue			
	Green			
				1000

Step 4. Estimate the population totals. If there were 1,000 cabs in the city, how many of them would be Blue? According to the story, 10% are Blue. That means 10 out of every 100 or 100 out of every 1,000 are Blue. That number,

100, is the left column total. The rest are Green. So 1,000 - 100 = 900 is the right column total. We put these column totals into the table:

Cabs in the City

Witness said:		Blue	Green	
	Blue			
	Green			
		100	900	1000

WARNING. The method we're teaching you for solving these problems won't work if you start out estimating the wrong totals. It's important in this step to correctly identify which part of the problem gives population information and which gives specific information that does not indicate any specific case. The specific information fingers a particular case.

Step 5. Fill in the cells. Working with each total, divide it among its two cells. First, for the 100 blue cabs, how many would the witness correctly see as Blue, and how many would the witness incorrectly see as Green? The story states that the witness is correct 70% of the time. So: 100 x .70 = 70 is the number of Blue cabs the witness would correctly call Blue, and the remaining, 100-70 = 30, are the number of Blue cabs the witness would incorrectly all Green.

Now consider the 900 Green cabs. Again the witness' accuracy is 70%: 900 x .70 = 630 is the number of Green cabs the witness would have correctly called Green. This number, 630, goes in the Green-Green cell. The rest of the Green cabs, 900 - 630 = 270, is the number of Green cabs the witness would have incorrectly called Blue.

Our table now looks like this:

Cabs in the City

Witness said:		Blue	Green	
	Blue	70	270	
	Green	30	630	
		100	900	1000

Comment. Notice that we now could, if we wished, find the last two totals, the total number of times the witness would have said "Blue", rightly or wrongly:

$$70 + 270 = 340$$

and the total number of times the witness would have said "Green", rightly or wrongly:

$$30 + 630 + 660$$

These totals are not intuitively obvious. The reason is that these totals are the total number of times the witness says "Green" and Blue". What the witness says depends not only on the witness' accuracy but also on the relative proportions of Blue and Green cabs the subject might have seen. You have to take both these facts into consideration to calculate the totals. In contrast, the population totals make a lot of sense, because they depend on only one kind of information, not two kinds. The total number of Blue cabs in the city is directly calculated as a percentage of the total number of cabs, regardless of what the witness might testify. This distinction is important because it shows you another way of telling, in any problem, which is the population information (that you start with in Step #4) and which is the specific information. The population information is information that directly translates into number totals. The specific information is information that does not translate into number totals because those number totals depend not only on the specific information but also on the population information.

In summary, here are two criteria (one discussed earlier) for telling which is which:

The population information:
(a) is general, background information and
(b) can be translated directly into number totals.
The specific information:
(a) specifies or identifies one case and
(b) cannot be directly translated into number totals because those totals also depend on the population information.

Step 6. Cross out the false. The witness in the story in fact testified that the cab was Blue. So the number of times the witness might have said "Green" is irrelevant to the problem. We cross out these false cells so we won't be tempted to use them in the next step:

		Cabs in the City	
		Blue	Green
Witness said:	Blue	70	270
	Green	~~30~~	~~630~~
		100	900

do not forget to cross

1000

<u>Step 7. Find the needed probability</u>. The two remaining cells are what we need to answer the question. They show that the witness would have said "Blue" <u>correctly</u> 70 times and would have said "Blue" <u>incorrectly</u> 270 times. From these two numbers we can get our probability.

If you're not used to thinking about probabilities, a nice way to think about them is to imagine that you fill an urn with 70 balls labeled "cab is really Blue" and 270 balls labeled "cab is really Green", for a total of 340 balls. Now sample one ball at random from the urn. What is the probability that the ball will be labeled "cab is really Blue?" the answer is the number of "cab is really Blue" balls divided by the total number of balls in the urn:

$$\frac{70}{70+270} = \frac{70}{340} = .21$$ (well, it;s really .2058...but we rounded it)

In other words, we <u>divide the number in the TARGET cell by the sum of the two numbers left in our table</u>. The TARGET cell is the one cell identified by both the specific information given in the problem ("a witness identified the cab as Blue") <u>and</u> the question asked at the end of the problem ("What is the probability that the cab involved in the accident was Blue?"). So the target cell is the "cab is Blue/Witness said Blue" cell.

That it. The answer, .21, is the probability that the hit-and-run cab was a Blue cab.

Are you surprised by the answer? Most people think that the correct answer should be .70, the same as the witness' accuracy. They tend to forget the population information, that is, they fail to notice that because there are so many more Green cabs than Blue cabs, there are also many more opportunities for the witness to be wrong when saying Blue.

Comment. While it's not necessary to solve the problem, it might help you to understand what's going on by thinking about this: What if the witness had testified that the cab was Green? Look back at the last table, the one with two crossed-out cells. Those crossed-out cells show 30 really Blue cabs and 630 really Green cabs. So the probability that the cab is really Green, if the witness said it was Green, is:

$$\frac{630}{630+30} = \frac{630}{660} = .95$$

This probability is higher than either the proportion of Green cabs in the city (90%) or the accuracy of the witness (70%). That's because in this case both pieces of information-- the population proportion and the witness' testimony, point in the same direction, towards Green.

Intermediate probabilities like .21are found only when the two pieces of information point in opposite directions: the witness said Blue but most cabs are Green.

That's the end of the tutorial. On the next page is a problem for you to do. Before doing the problem:

1. Review the tutorial to make sure you understand it.
2. Ask any questions you have.

When you are ready, proceed to the problem on the next page. We are interested in how effective the tutorial is in teaching you how to do such problems. So while you are doing the problem, feel free to:

1. Review the tutorial again.
2. Use a hand calculator.
3. Ask question.

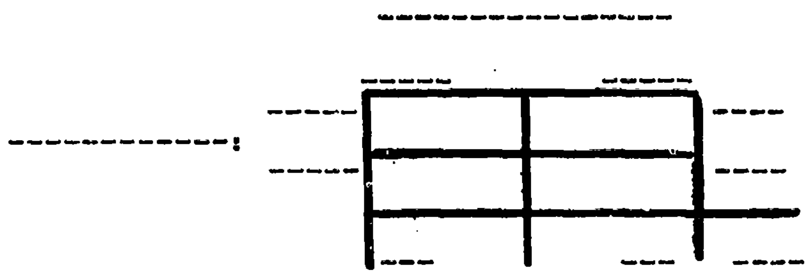
Please work the following problem using the method just described. We've drawn you a table to work with

A light bulb factory uses a scanning device which is supposed to put a mark on each defective bulb it spots in the assemble line. Eighty five percent (85%) of the light bulbs on the line are OK; the remaining 15% are defective.

The scanning device is known to be accurate in 80% of the decisions, regardless of whether the bulb is actually ok or actually defective. That is, when a bulb is good, the scanner correctly identifies it as good 80% of the time. When a bulb is defective, the scanner correctly marks it as defective 80% of the time.

Suppose someone selects one of the light bulbs from the line at random and gives it to the scanner. The scanner marks this bulb as defective.

What is the probability that this bulb is really defective?



Step 1. Draw a table. Done.

Step 2. label the table.

Step 3. Assign an arbitrary grand total. Use 1,000.

Step 4. Estimate the population totals. First decide which set of information is population information. Then divide the 1,000 into two parts, using information from the problem.

Step 5. Fill in the cells. Divide each of your estimated totals among its two cells, according to the information in the problem.

Step 6. Cross out the false. Cross out the two cells that are contradicted by the information given in the problem.

Step 7. Find the needed probability. Write the relevant numbers in the top and bottom of the fraction and convert the fraction to a decimal answer.

# in target cell
______________________ = ____________ = ______ = . , answer.
Sum of #'s in both cells

Algorithm for the light bulb problem

A. Out of 1,000 light bulbs produced by the factory, How many are defective? Multiply the percentage of defective bulbs by 1,000. (First convert the percentage value to a decimal value before multiplying).

1,000 x _______________ = _______ (A)
Proportion of
Defective Bulbs

B. Subtract you estimate in (A) from 1,000 to get the number of bulbs out of 1,000 that are NOT defective.

1,000 - (A)_______________ = _______ (B)

C. What percentage of the time is the scanner able to correctly identify light bulbs that are actually defective? (from the problem) ____________ (C)

D. What percentage of the time is the scanner able to correctly identify light bulbs that are actually <u>not</u> defective? (from the problem) ____________ (D)

E. Look over the following table:

LIGHT BULBS ARE:

	Actually Defective	Not Defective	
Scanner Say is Defective	Box # 1	Box # 4	_____
Scanner Says is NOT Defective	Box # 2	Box # 3	
	_______ (A)	+ _______ (B)	

F. Write the number of defective light bulbs from (A) on the line labeled (a) in the table above, just below Box # 2.

G. Write the number of non-defective light bulbs from (b) on the line labeled (b) in the table above, just below Box # 3.

H. Multiply the percentage value in (c) by your estimate from (A). (First convert the percentage value to a decimal value before multiplying).

(A)________ x (C)________ = _______ (H)

Write you value for (H) in Box # 1.

I. Subtract you value in (H) from you value in (A).

(A)________ - (H)________ = _______ (I)

Write you value for (I) in Box # 2.

J. Multiply the percentage value in (D) by your estimate from (B). (First convert the percentage value to a decimal value before multiplying).

(B)________ x (D)________ = _______ (J)

Write your value for (J) in Box # 3.

K. Subtract your value in (J) from your value in (B).

(B)________ - (J)________ = _______ (K)

Write you value for (K) in Box # 4.

L. Add the numbers in Boxes # 1 and # 4.

Box # 1________ + Box # 4________ = _______ (L)

Write you value for (L) on the line labeled (L), to the right of the boxes.

M. To get the final answer, divide your value in Box #1 by your value for (L).

Box # 1________ ÷ (L)________ = _______ (M)

APPENDIX F

Training Problems

The Parachute Problem

A factory that manifactures parachute-brakes for aircrafts, uses a special device for checking the parachutes.

Eighty five percent of the parachutes are in order; the other 15% are defective.

The device is known to be accurtate in 70% of the cases. That is, 70% of the parachute that are in order and 70% of defective parachues will be correctly identified as such.

A parachute was randomly selected from the line, and was checked by the device. The device identified it as defective.

What are the chances that the parachute is really defective?

The Jaundice Problem

Jaundice is a disease that may occur in two different forms: viral and infectious.

Twenty percent of jaundice cases are infectious; the other 80% are viral. The symptoms of the two forms are identical, but the treatment is different. Inadequate treatment may cause severe side effects, and therefore the type of jaundice in each case must be identified correctly.

A certain blood test is used to distinguish between the two forms of jaundice. The test results are known to be accurate in 80% of the cases. That is, both infectious and viral jaundice will be correctly idenfitied as such in 80% of the cases.

A soldier who had symptoms of jaundice was administered this blood test. The test results indicated infectious jaundice.

What are the chances that the soldier really had Infectious Jaundice?

Generalization Problems

The Missile Problem

Intelligence sources revealed that a hostile Army had purchased sophisticated G-7 anti aircraft missiles.

Israeli industry had developed a special device, capable of receiving the signals broadcasted from anti aircraft missiles that enabled identification of missile type. The device is known to be accurate in 80% of the cases, that is, a G-7 missile type and misslies of "other types" will be correctly identified as such in 80% of the cases. Knowledge of the exact type of missile, improves the defence profile an aircraft flying in bound area.

Researches done by the Tactical Warfare Development Committee show that the chances for launching a G-7 type missile is 10%.

Neutral risk - An anti aircraft missile has been sent to a certain area, and was identified by the device as being of type G-7.

General risk - An anti aircraft missile has been launched to a certain area where only Israeli aircraft fly, and was identified by the device as being of type G-7.

Personal risk - Suppose you are a pilot flying an Israeli aircraft. An anti aircraft missile that had been launched to the area where you are flying was identified by the device as being of type G-7.

What are the chances that this missile is really a type G-7 missile?

The Masks Problem

Due to failures in production 20% of NBC (Nuclear biological & Chemical) Masks are defective; the other 80% are in order.

A defective mask can be identified by using a simple device. This device is known to be accurate in 95% of the cases. That is a defective mask and a mask which is in order will be correctly identified as such in 95% of the cases.

Neutral risk - A mask was identified as defective.

General risk - A mask, which is part of the personal equipment of a certain soldier, was identified as defective.

Personal risk - Due to warning of potential NBC attack, soldiers were given personal masks. Your mask was identified as defective.

What are the chances that the mask is really defective?

<u>The Seals problem</u>

Ten percent of brake-cylinder seals, used as spare parts for armoured vehicles are defective.

A special device is used to check the seals. The device is known to be accurate in 95% of the cases. That is to say, a good seal and a defective seal will be correctly identified as such in 95% of the cases.

Neutral risk - A seal was checked and was found to be in order.

General risk - A seal was checked and was found to be in order. The seal was instaled in an armoured vehicle which was sent on a dangerous mission.

Personal risk - A seal was checked and was found to be in order. The seal was instaled in an armoured vehicle in which your are a crew-member, and this vehicle was sent on a dangerous mission.

<u>What are the chances that the seal is really in order</u>?

APPENDIX G

Training by Mental Image (TbMI)

Consider the following problem:

8% of the male population is color blind.

A test of color blindness is known to be accurate in 90% of the cases, that is, 90% of color blind men will be correctly identified by the test as color blind. Of those who are not color blind, 90% will be correctly identified as having normal color perception.

A certain man was classified, according to test results, as color blind.

What are the chances that this man is really color blind?

Answer ______________.

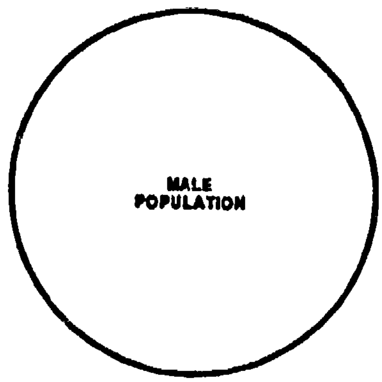
Research has shown that, when presented with such problemsprobability estimates tha vary as follows: 90%; 8%; 7.2%; 41%.

The reason for such a variety of answers is that people understand the problem and its solution, in different ways, not all of which are correct.

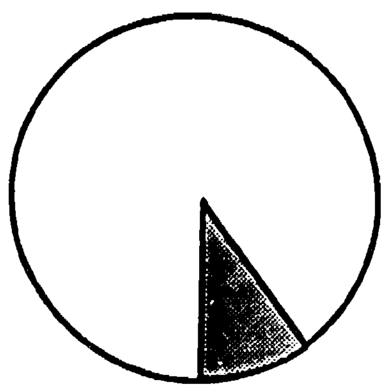
In the following pages we present a tutorial to help you solve such problems correctly. Please read through the tutorial carefully, and solve the problems that will be presented to you, according to the instructions.

Tutorial

The population discussed in the previous problem is the male population. Suppose the circle we have drawn represent this population.



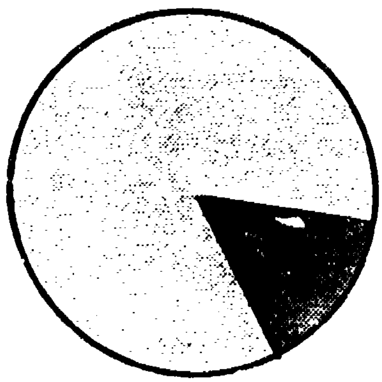
We already know that 8% of this population is color blind. The following drawing illustrates the division of the overall population (the male population) into two sub populations (those who are color blind and those who have normal color vision).



Area [shaded box] represents the men who are color blind;

Area [white box] represents the men who have notmal color vision.

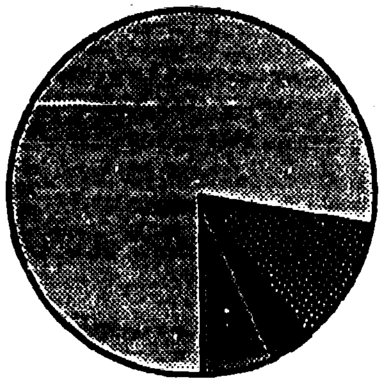
The problem's storyline states that "A test of color blindness is known to be accurate in 90% of the cases". If the entire male population were tested, it would be possible to divide the population into two groups of people. One group would contain those men who would be identified as color blind and the other would contain those men who would be identified as having normal color vision. But, since the test is only partially accurate, division of the population according to test results would not represent the actual "state-of-the-world". The division according to <u>test results</u> is illustrated in the following drawing.



Area  represents the men that would be identified by the test as color blind (this area was originally red);

Area  represents the men that would be identified by the test as having normal color vision (this area was originally yellow

To illustrate the inaccuracy of the test we can superimpose the two drawings.



Area represents the color blind men that would be identified by the test as color blind;

Area represents color blind men that would be identified by the test as having normal color vision;

Area represents men having normal color vision that would be identified by the test as color blind;

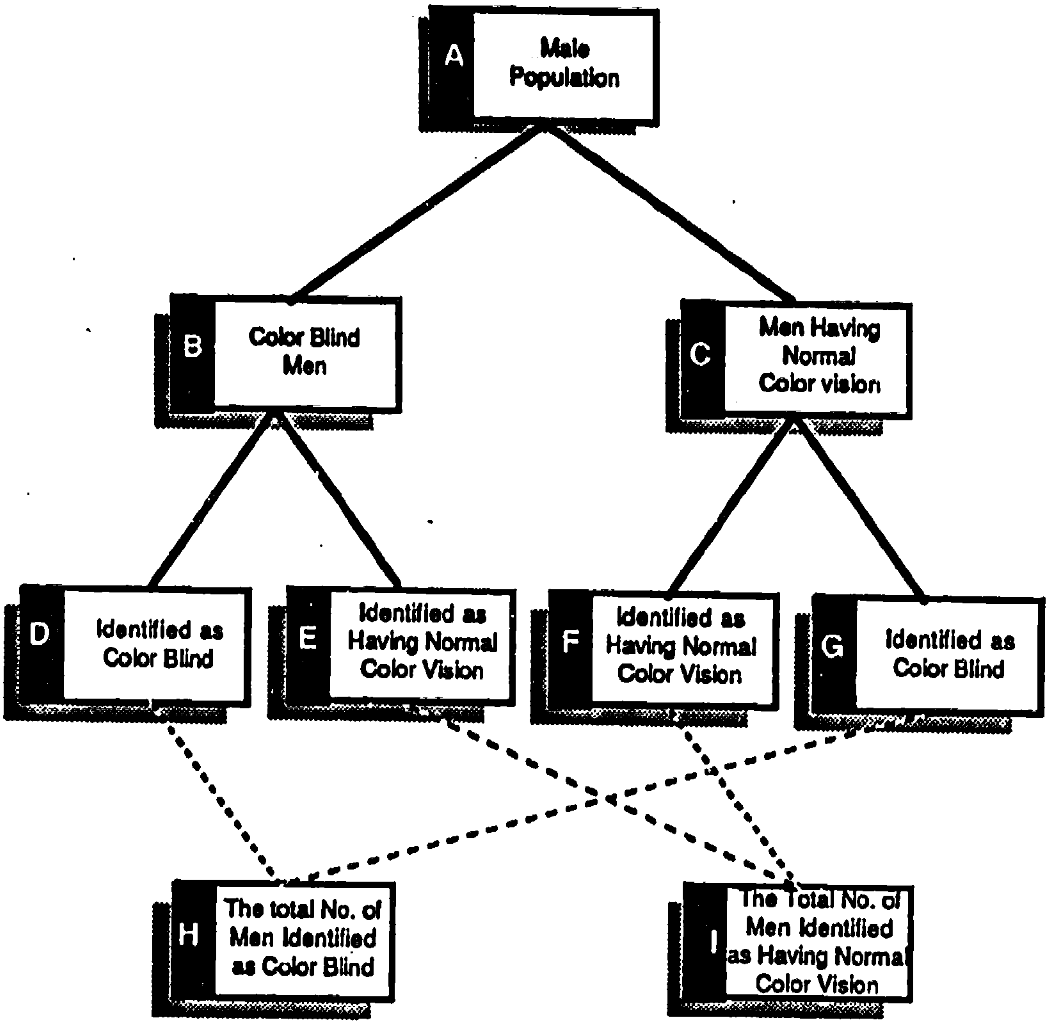
Area represents men having normal color vision that would be identified by the test as having normal color vision.

Note that in some of the cases the test results are in error, that is, they do not reveal reality. In fact there are two types of errors:

<u>"Miss"</u>: A color blind man, who is identified by the test as having normal color vision.

"<u>False Alarm</u>": A man having normal color vision, who is identified by the test as color blind.

An alternative way to present the total population and its division to the various sub population is by using a "tree". Look through the "tree" carefully.


A
Male Population
B
Color Blind Men
C
Men Having Normal Color vision
D
Identified as Color Blind
E
Identified as Having Normal Color Vision
F
Identified as Having Normal Color Vision
G
Identified as Color Blind
H
The total No. of Men Identified as Color Blind
I
The Total No. of Men Identified as Having Normal Color Vision

The problem's storyline also state: "A certain man was classified, according to test results, as color blind". This means that this man is a member of the sub population entitled "total identified as color blind" in the "tree".

We already know, that this sub population ("identified as color blind"), includes two kinds of men:

a. Color blind men who were identified by the test as color blind;

b. Men having normal color vision who were identified by the test as color blind (false alarm).

We are requested to indicate" <u>What are the chances that this man is really color blind</u>?". In other words, out of all the men who were identified by the test as color blind, what is the actual proportion of color blind men? The appropriate calculation is:

$$\frac{\text{Color Blind Men Identified as Such}}{\text{The Total No. of Men Identified as Color blind}} = \text{The Chances that the Man is Really Color blind}$$

In order to calculate this, we have to determine the appropriate values for each the sub-population represented in the "tree" and the drawing. This can be done by using data extracted from the problem itself, and according to the following instructions.

1. Suppose the male population contains 1000 men. This value will be written In the "tree" on page 9, in the frame marked A.

2. Of these 1000 men, how many are really color blind?

$$\underset{\text{The total population}}{1000} \times \underset{\text{The Proportion of Color Blind Men}}{.8} = \underset{\text{The Total No. of Color Blind Men}}{80}$$

This value will be written In the "tree" on page 9, in the frame marked B.

3. Of these 1000 men, how many have normal color vision?

1000	-	80	=	920
The total population		The Total no. of Man having Normal Color Vision		The Total No. of Color Blind

The resulting value will be written In the "tree" on page 9, in the frame marked C.

4. Of all color blind men, how many will be identified as color blind?

80	X	.90	=	72
The Total No. of Color Blind		The accuracy of test results		The Total No. of Color Blind Men Identified as Color Blind

The resulting value will be written In the "tree" on page 9, in the frame marked D.

5. Of all color blind men, how many will be identified as having normal color vision?

80	-	72	=	8
The Total No. of Color Blind men		The Total No. of Color Blind Men Identified as Color Blind		Color Blind Identified as Having Normal Color Vision

The resulting value will be written In the "tree" on page 9, in the frame marked E.

6. Of all the men having normal color vision, how many will be identified as having normal color vision?

920	X	.90	=	828
The Total No. of Man Having Normal Color Vision		The accuracy of test results		men having Normal color vision Identified as such

The resulting value will be written In the "tree" on page 9, in the frame marked F.

7. Of all men having normal color vision, how many will be identified as color blind?

920	-	828	=	92
The Total No. of Man Having Normal Color Vision		Man Having Normal Color Vision Identified as Such		Man Having Normal Color Vision Identified as Color Blind

The resulting value will be written in the "tree" on page 9, in the frame marked G.

8. What is the total of men identified color blind?

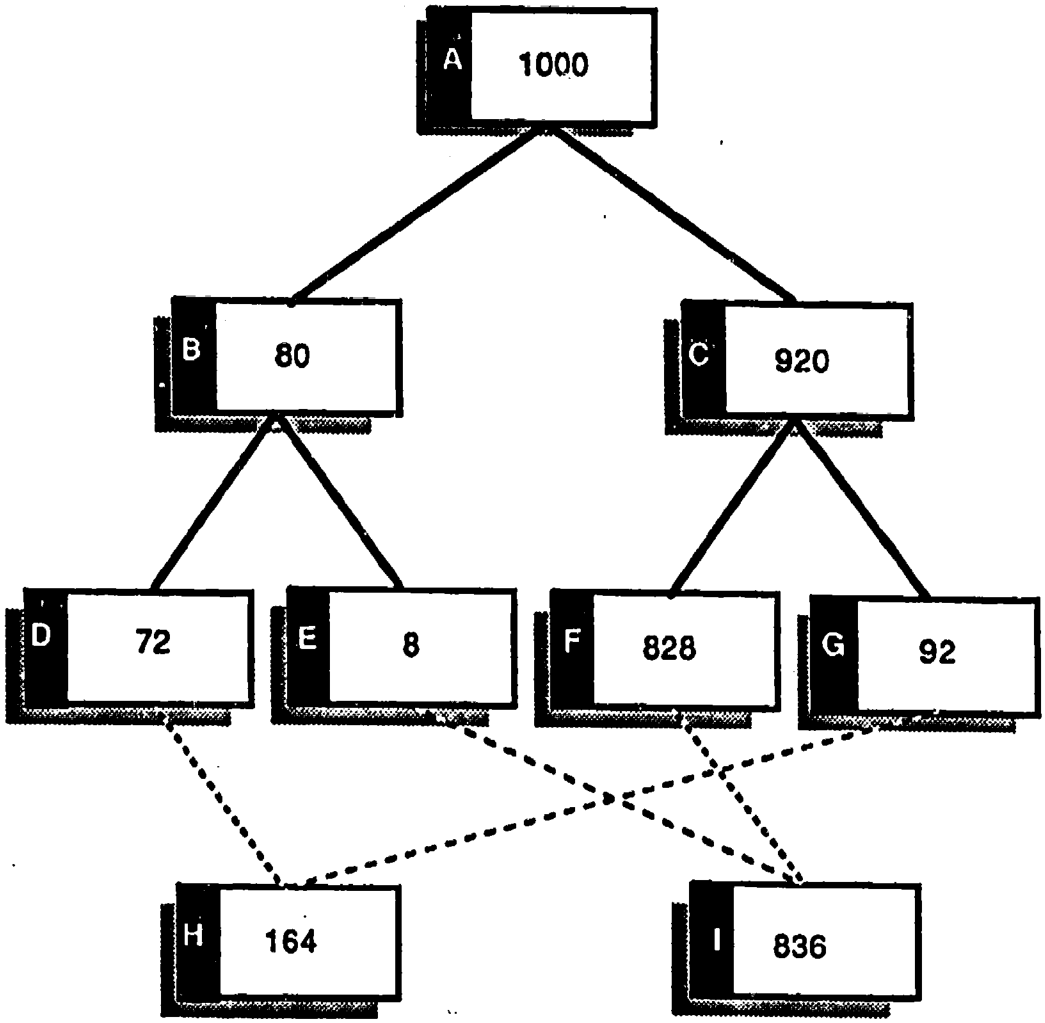
72	+	92	=	164
The Total No. of Color Blind Men Identified as Color Blind		Man Having Normal Color Vision Identified as Color Blind		The Total No. of Men Identified as Color Blind

The resulting value will be written in the "tree" on page 9, in the frame marked H.

9. What is the total of men identified as having normal color vision?

828	+	8	=	836
Man Having Normal Color Vision Identified as Such		Color Blind Identified as Having Normal Color Vision		The Total No. of Men Identified as Having Normal Color Vision

The resulting value will be written in the "tree" on page 9, in the frame marked I.


A 1000
B 80
C 920
D 72
E 8
F 828
G 92
H 164
I 836

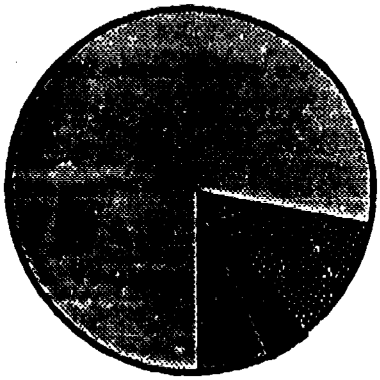
Now we have all the data required to calculate the target probability. The calculation will be as follows:

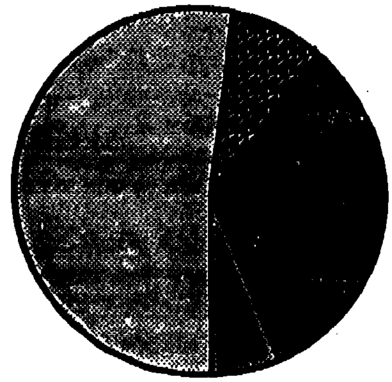
72	:	164	=	0.41
Color Blind Men Identigled as Such		The Total No. of Men Identified as Color Blind		The Chances that the Man is Really Color Blind

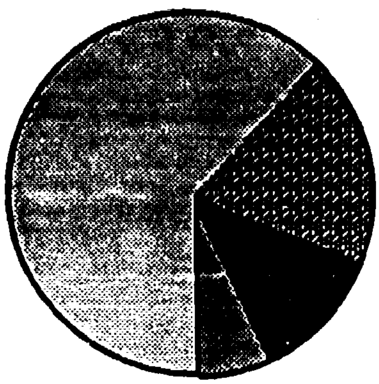
Notice! The correct answer is 0.41, that is, 41% chance that a man who was identified according to test results as color blind, is really color blind.

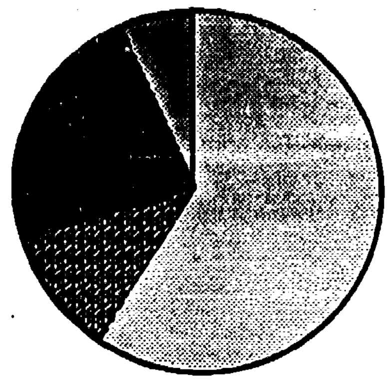
This answer may seem unreasonable to you. In this case, you are advised to go over the tutorial again. You may be able to understand the situation discussed in the problem better, and the method of solution, if you think about how the final results (i.e., the chances that a man who was identified according to test results as color blind really being color blind) would change, if the percentage of color blind men in the overall population was different.

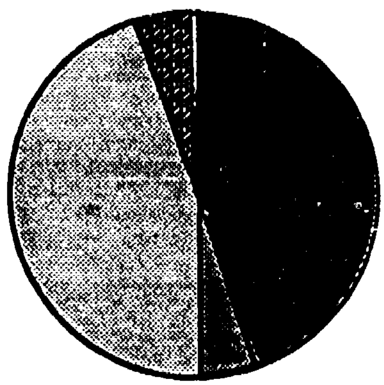
This will be illustrated by drawings. Each one of the following drawings represents a population in which the percentage of color blind men [illegible] is different. The degree of accuracy of test results remains constant.

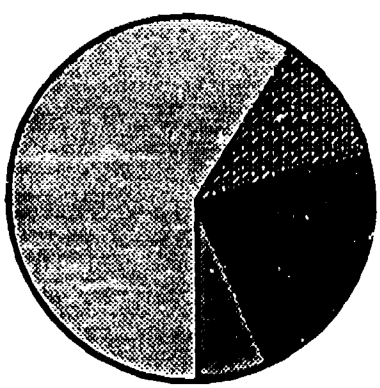


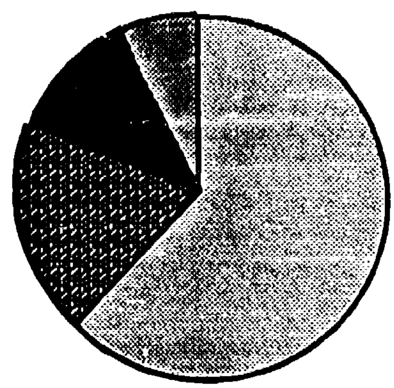


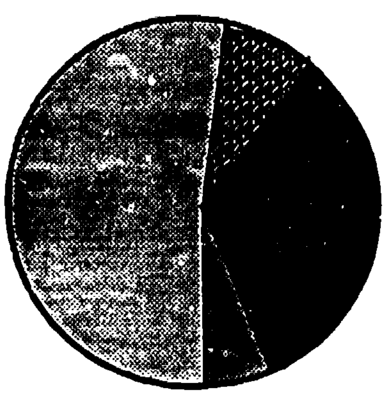


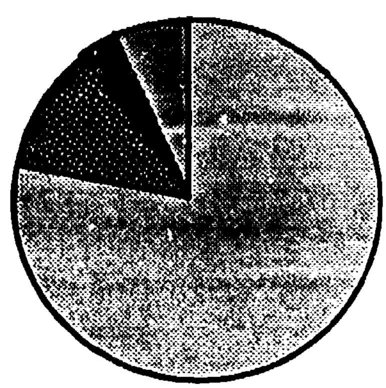












Notice! In each drawing, you can note the size of the area marked [redacted] relative to the area marked [redacted] This is the proportion of color blind men who are, in fact, color blinde and who were identified as such. In addition you can note the size of the error - misses + false alarms - of the test results.

The situations discussed here, contain two "rules" used to determine the target probability. These "rules" are:

1. The distribution of the phenomenon in the population.

2. The degree of test accuracy (to what degree they reveal reality).

Changing the value of either "rule" will change the size of the various errors, and therefor the target probability.

A special case is that in which the percentage of color blind men in the population is 50. Here, the phenomenon of color blindness is distributed randomly. That is, the phenomenon is not distributed according to any "rule", and therefore this information has no influence. In this case, the only relevant information is the percentage of cases in which the test results are accurate (in our story, 90%).

This is the end of the tutorial. In the following pages you will be presented with additional problems. You are requested to solve these problems according to the tutorial. Before continuing you are advised:

1. To read the tutorial again, and make sure you understand it.

2. If you have any questions, you may consult the experimenter.

When your are ready, go on to the next page. while answering the next two problems, you may:

1. Read the tutorial again.

2. Use a calculator.

3. Ask questions.

The data in Table 33 show that on the average subjects are generally confident in the accuracy of their answers. The mean confidence rating across all the data is 5.60. The mean ratings are higher for the A groups than the UA groups. The mean ratings are also higher for the NTL groups than the TL groups. Of the various risk level groups, the mean confidence rating for NR groups is the highest, followed by the PR groups and then the GR groups. An analysis of variance performed on these data showed significant main effect for the aid condition ($F(1,151)=8.42$, $p<.01$).

Reasonableness

The Reasonableness judgement were found to be unaffected by the aid, time limit and risk conditions. The number of "yes" responses for all three generalization problems for each subject was computed. An analysis of variance performed on these data failed to reach significance.

Subjective Mental Load

The mean ratings and s. d. for difficulty, mental effort, fatigue, frustration, subjective time stress and the subjective mental load measure, are shown in Tables 34 to 38.

Table 34: Mean and s.d. of Difficulty Rating (Upper No.= Mean; Lower No.=s.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	2.98 1.78	2.75 1.75	2.85 1.76
LIMITED	3.20 1.55	2.91 1.89	3.05 1.76
TOTAL	3.09 1.66	2.83 1.81	2.95 1.74

Table 35: Mean and s.d. of Mental Effort Rating (Upper No.= Mean; Lower No.=s.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	2.91 1.40	3.27 1.60	3.11 1.51
LIMITED	3.60 1.54	2.79 1.35	3.18 1.49
TOTAL	3.27 1.50	3.04 1.50	3.15 1.50

Table 36: Mean and s.d. of Fatigue Rating (Upper No.= Mean; Lower No.=s.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	3.32 1.93	2.77 1.81	3.02 1.88
LIMITED	3.36 1.69	2.23 1.17	2.78 1.55
TOTAL	3.34 1.80	2.50 1.55	2.90 1.72

Table 37: Mean and s.d. of Frustation Rating (Upper No.= Mean; Lower No.=s.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	3.04 2.14	3.38 1.89	3.22 2.00
LIMITED	3.14 1.83	3.27 2.00	3.21 1.91
TOTAL	3.09 1.97	3.32 1.94	3.21 1.95

Table 38: Mean and s.d. of Subjective Mental Load (Upper No.=Mean; Lower No.=s.d.)

TIME \ AID	AIDED	UNAIDED	TOTAL
UNLIMITED	2.81 1.56	2.72 1.99	2.77 1.75
LIMITED	3.12 1.57	2.55 1.62	2.84 1.77
TOTAL	3.01 1.66	2.64 1.49	2.82 1.85

Tables 34 to 38 indicate that the mean ratings of difficulty, mental effort, fatigue and the computed subjective mental load measure were higher for the TL groups than the NTL groups, and higher for the A groups than the UA groups, the ratings are also higher for NR groups followed by GR groups and the PR groups, in that order. The mean ratings for frustration were similar to the above for time limit and risk, but higher for the UA groups than the A groups.

The mean ratings were similar for difficulty, mental effort and fatigue ratings for aid and risk, but higher for the TL groups than the NTL groups.

An analysis of variance on these data indicated significant effect of time limit ($F(1,192)=7.88$, $p<.01$) and aid ($F(1,192)=17.76$, $p<.01$) on subjective time stress only. Other main effect failed to reach significance. The two-way interaction of time limit x aid was found to be significant for the mental effort ($F(1,192)=2.56$, $p<.05$), fatigue ($F(1,192)=13.31$, $p<.01$) and time stress ($F(1,192)=5.06$, $p<.05$) dimensions. The two-way interaction for time limit x aid was found to be significant for the effort dimention ($F(1,192)=7.88$), $p<.01$). The two-way interaction for time limit x risk leve was found to be significan for the fatigue ($F(1,192)=13.31$, $p<.05$) and subjective time stress ($F(1,192)=4.41$, $p<.05$) dimensions. The interactions are shown in Figures 3 to 5.

Figure 3: Mean Mental Effort Ratings as Function of Time Restriction and Aiding Conditions

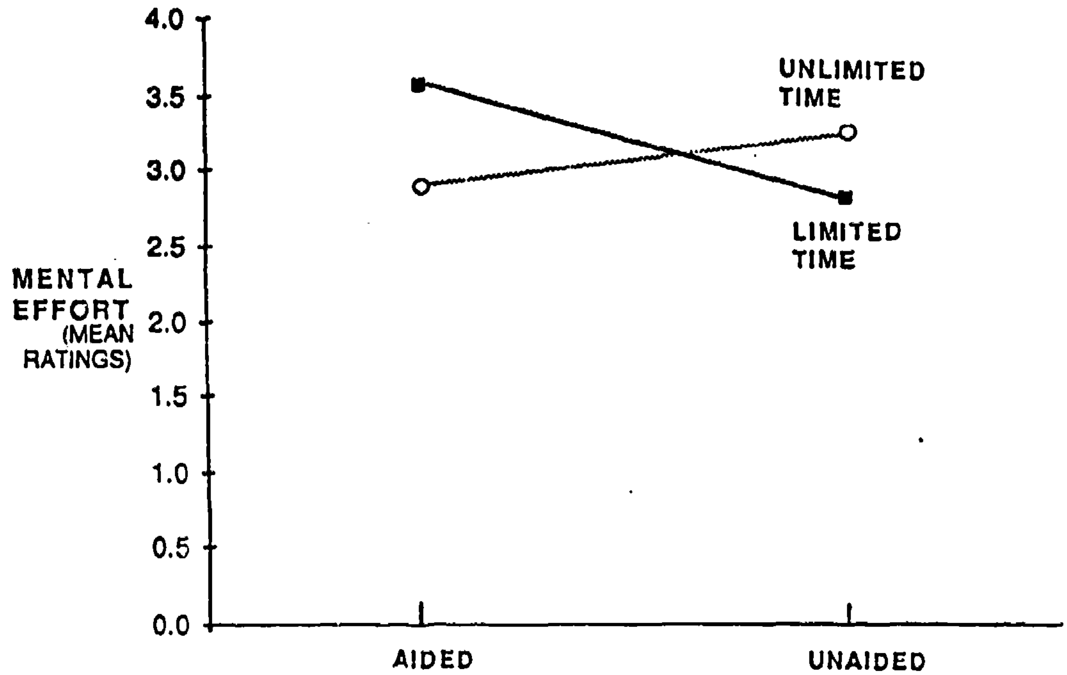


Figure 4: Mean Fatique Ratings as Function of Time Restriction and Risk Conditions

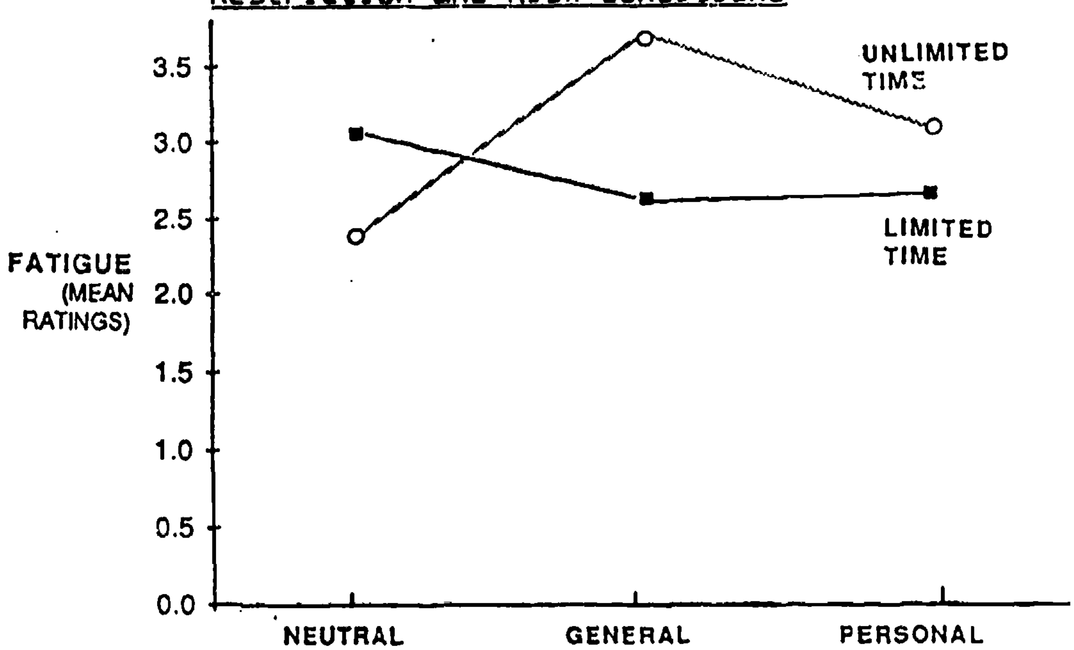
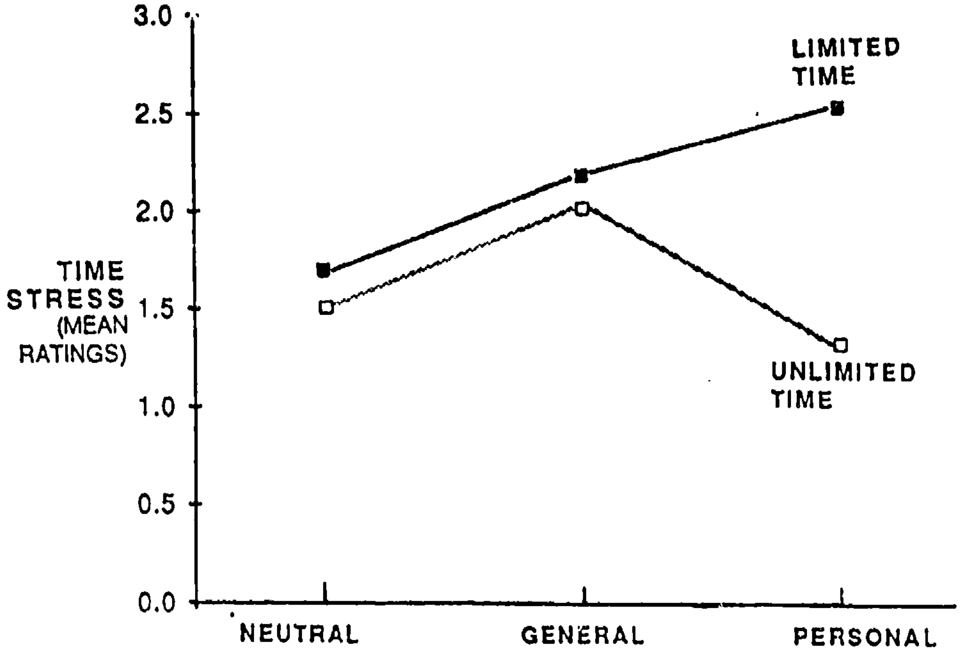


Figure 5: Mean Time Stress Ratings as Function of Time Restriction and Risk Conditions



Simple tutorial Vs. TbMI

The results of Experiment III, obtained with the aid of the tutorial, were compared with those obtained with the TbMI aid under unlimited time conditions and neutral risk, when performing the missile problem in Experiment IV. This group was selected since it represents experimental conditions similar to those of Experiment III. The Missile problem was selected since it was the first generalization problem presented to the subjects. Subjects' responses to the compared generalization problems are shown in Table 39. The data in Table 39 show that the proportion of "correct" responses was higher for subjects in Experiment IV (72%) than for subjects in Experiment III (55%).

Table 39: Frequencies and Proportions of Subjects' response Mode when Aided by Simple Tutorial and TbMI

RESPONSE \ AID TYPE	TUTORIAL	TbMI
CORRECT	16 55.2%	13 72.2%
DIAGNOSTIC	3 10.3%	1 5.6%
BASE RATE	3 10.3%	0 0%
CONDITIONAL	1 3.4%	0 0%
OTHER	6 20.7%	4 22.2%

While no significant difference, between the two response distributions, was shown by using a chi-square test, Table 39 clearly shows an overall trend, which indicates the relative advantage of the TbMI method.

Discussion of Part B

The major finding of experiment IV, is that the use of mental images, for presentation and organization of the verbal explanation in the Training by Mental Image TbMI, contributed considerably to the effectiveness of this aid, under all the experimental conditions. The results indicate that the TbMI method succeeded in improving performance, creating constructive change in the conceptualization of base-rate problems, and in acquiring new cognitive skills with which to examine these problems.

The way of solution specified in the tutorial used by Lichtenstein & MacGregor, (1985) and in Experiment III, was similar to the one specified in the TbMI. This enabled comparison between the two tutorials that would reveal the contribution of the mental images as a way of presentation. This comparison showed that the TbMI led to a better generalization than the original tutorial.

The risk manipulation did not influence performance. This may indicate that the risk element in the generalization problems did not affect the interpretation and judgements of relevance of the different information types. An alternative explanation is that, as hypothesized, the cognitive skills acquired through the TbMI method, were strong enough to overcome the risks influence. This is also supported by the interaction effect of time limit x risk for the confidence ratings. That is, if risk had no influence at all, the confidence rating would not be influenced, but since they were, it means that this influence was removed after training.

The time limit manipulation had a minor effect on subjects performance. This also indicates of the effectiveness of the aid and its generalized effect, especially in view of previous findings that framing is not transferred to stress condition (Zakay, 1984), and in view of the validation of time restriction manipulation in Experiment IV. However, this manipulation did affect confidence ratings for one generalization problem. This may indicate that although the training method was effective, more training is required, in order to make the new way of conceptualization more intuitive.

The degree of confidence, in the accuracy of the answers, to two of the generalization problems, was influenced by the time restriction and aiding manipulations. Trained subjects reported higher confidence then untrained subjects. The

subjects who had to solve the generalization problems under time limit conditions, reported lower confidence than those who were not time restricted. This indicates that the training method decreased confusion and uncertainty with which the untrained subjects had to deal. However, more training may be required for time stress conditions.

The reasonableness judgements were not affected by the time restriction and the training manipulations. This is surprising, in light of their influence on confidence ratings. This may be the result of cognitive dissonance.

The subjective mental load measures were affected by aiding and time limit. Higher subjective mental load was reported when subjects performed under time limit condition, and when presented with the training method. Similar finding was reported by Einhorn, (1970), who found that using heuristics require less effort. It should be noted, that the subjects answered the subjective mental load questionnaire after completing the training and the generalization problems. Although it was emphasized that the questionnaire related only to the generalization problems, it is likely that the training had an effect on these measures.

The use of mental images for presentation and organization of verbal material, can be applied in developing Computer Aided Instruction (CAI). An attempt in this direction has already been made. Preliminary program for interactive learning, using IBM-XT computer, was developed. This program focuses only on training for base-rate problems solution, according to the same method used in the TbMI. Pretesting has revealed that this concept is promising, but needs further developement.

GENERAL CONCLUSIONS

A. Regarding the algorithmic decomposition aid for estimating unknown quantities, the results suggest that in developing an aid or training method, based on the algorithmic approach, the unique characteristics of the target population, should be taken into account. The aid and the training method must be adjusted accordingly, in order to be compatible with the thinking patterns and cognitive style of the target population. Only after the aid and training method, are adapted to the population, the members can compose individual algorithms to match its content and organization to their own cognitive style and thinking patterns.

B. The TbMI method in solving base-rate problems is effective and led to systematic change of the way in which people conceptualize and solve base-rate problems. This method should be further developed.

C. The use of mental images for presentation and organization of verbal material, can be applied in developing Computer Aided Instruction (CAI).

D. In view of the success of the TbMI method, in solving base-rate problems, it is recommended to apply this approach as a training method for general estimation problems.

REFFERENCES

Adams, J., A., 1976, Learning and Memory: an Introduction, The Dorsey Press, Homewood, Illinois.

Ben Zur, H., & Breznitz, S. J., 1981, The Effect of Time Pressure on Risky Choice Behavior. Acta Psychologica 47, 89-104.

Beyth-Marom, R., & Fischhoff, B.,1983, Attitudes and Social Cognition: Diagnosticity and Pseudodiagnosticity, Journal of Personality and Social Psychology, 45, 6, 1185-1195.

Beyth-Marom, R., Dekel, S., Gombo, R., & Shaked, M., 1985, An Elementary Aproach to Thinking Under Uncertainty, Lawrence Erlbaum Associates, Inc.Hillsdale, New Jersey.

Casscells, W., Schoenberger, A., & Grayboys, T. B., 1978, Interpretation by physicians of clinical laboratoy results. New England Journal of Medicine, 299, 955-1000.

Eddy, D. M., 1982, Probabilistic Reasoning in Clinical Medicine: Problems and Apportunities. In D. Kahneman, P. Slovic and A. Tversky, Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press.

Edwards, W., 1968, Conservatism in Human Information Processing. In B. Kleinmunts (Ed.), Formal Representation of Human Judgment, New York: Wiley.

Einhorn, H. J., 1970. The Use of Nonlinear Noncompensatory Models in Decision Making. Psychological Bulletin 73, 221-230.

Fischhoff, B., Slovic, P., & Lichtenstein, S., 1979, Subjective Sensitivity Analysis. Organizational Behavior and Human Performance 23, 339-359.

Fischhoff, B., & Bar-Hillel, M., 1984, Focusing Techniques: A Shortcut to Improving Probability Judgments? Organizational Behavior and Juman Performance 34, 175-194.

Hammerton, M. A., A Case of Radical Probability Estimation. Journal of Experimental Psychology 101, 252-262.

Kahneman, D., & Tbersky, A., 1972a, On predictic:: and judgment. ORI Research Monograph, 12(4).

Kahneman, D., & Tversky, A., 1973, On the Psychology of Prediction. Psychological Review 80, 237-251.

Kahneman, D., & Tversky, A., 1979, Intuitive Predictions: Biases and Corrective Procedures, TIMS Studies in Management Science, 12, 313-327.

Kahneman, D., Slovic, P., & Tversky, A., 1982, Judgment under Uncertainty: Heuristics and Biases, New York: Cambridge University Press.

Koriat, A., Lichtenstein, S., & Fischhoff, B., 1980, Reasons for Confidence. Jurnal of Experimental Psychology: Human Learning and Memory 6, 107-118.

Lichtenstein, S., & MacGregor, D., 1985, Structuring as an Aid to Performance in Base-Rates Problems, Decision Reasearch Report.

Lyon, D., & Slovic, P., 1976, Dominance of Accuracy Information and Neglect of Base Rates in Probability Estimation. Acta Psychologica 40, 287-298.

Nisbett, R. E., & Ross, 1980, L. Human Inference: Strategies and Shortcoming of Social Judgment. Englewood Cliffs, N.J.: Prentice-Hall.

MacGregor, D., Lichtenstein, S., & Slovic, P., 1985, Structuring Knowledge Retrieval, Decision Research Repost.

Paivio, A., 1971, Imagery and Verbal Processes. New York: Holt, Rinehart and Winston.

Tverskey, A., & Kahneman, D., 1973, Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psycholoty 5, 207-232.

Tverskey, A., & Kahneman, D., 1974, Judgment under Uncertainty: Heuristics and Biases, Science 185, 1124-1131.

Zakay, D., & Wooler, S., 1984, Time Pressure Training and Decision Effectiveness. Ergonomics, 27, 273-284

Zakay, D., 1985, Post-Decisional Confidence and Conflict Experienced in a Choice Process, Acta Psychologica, 58, 75-80.